JASA ARTICLE

The importance of processing resolution in "ideal time-frequency segregation" of masked speech and the implications for predicting speech intelligibility[a)]

Christopher Conroy,[b),c)] Virginia Best,[c),d)] Todd R. Jennings,[c)] and Gerald Kidd, Jr.[c)]
Department of Speech, Language and Hearing Sciences, Boston University, 635 Commonwealth Avenue, Boston, Massachusetts 02215, USA

ABSTRACT:
Ideal time-frequency segregation (ITFS) is a signal processing technique that may be used to estimate the energetic and informational components of speech-on-speech masking. A core assumption of ITFS is that it roughly emulates the effects of energetic masking (EM) in a speech mixture. Thus, when speech identification thresholds are measured for ITFS-processed stimuli and compared to thresholds for unprocessed stimuli, the difference can be attributed to informational masking (IM). Interpreting this difference as a direct metric of IM, however, is complicated by the fine time-frequency (T-F) resolution typically used during ITFS, which may yield target "glimpses" that are too narrow/brief to be resolved by the ear in the mixture. Estimates of IM, therefore, may be inflated because the full effects of EM are not accounted for. Here, T-F resolution was varied during ITFS to determine if/how estimates of IM depend on processing resolution. Speech identification thresholds were measured for speech and noise maskers after ITFS. Reduced frequency resolution yielded poorer thresholds for both masker types. Reduced temporal resolution did so for noise maskers only. Results suggest that processing resolution strongly influences estimates of IM and implies that current approaches to predicting masked speech intelligibility should be modified to account for IM.
© *2020 Acoustical Society of America*. https://doi.org/10.1121/10.0000893
(Received 18 July 2019; revised 28 January 2020; accepted 24 February 2020; published online 16 March 2020)
[Editor: Jennifer J. Lentz] Pages: 1648–1660

I. INTRODUCTION

"Ideal time-frequency segregation" (ITFS) is a signal processing technique that has been used in recent years to estimate the relative contributions of two different types of masking in speech-on-speech (SOS) masking experiments (e.g., Brungart *et al.*, 2006, 2009; Kidd *et al.*, 2016; Kidd *et al.*, 2019; Buss *et al.*, 2017; Rennies *et al.*, 2019). The first type, termed "energetic masking" (EM), occurs when a message of interest (the "target") overlaps with another (the "masker") in a given time-frequency (T-F) region and the masker is more intense, limiting the availability of target information in that region. Because the acoustic overlap of the two waveforms implies a corresponding overlap in the representation of the stimulus at the auditory periphery, EM is defined as masking that results from interactions between the target and masker at this stage of neural processing. "Informational masking" (IM), by contrast, is defined as masking that results from interactions between the target and masker at higher levels of the auditory system beyond the periphery (e.g., Lutfi, 1990; Durlach *et al.*, 2003; Watson, 2005; Kidd *et al.*, 2008a; Kidd and Colburn, 2017). IM may occur when the observer is uncertain or confused about which T-F regions of the target + masker mixture belong to the target as opposed to the masker, with the consequence being an inability to effectively segregate the target from the masker and/or selectively attend to the target over time.

ITFS seeks to emulate the effects of EM computationally by eliminating T-F regions of the target + masker mixture in which the masker is more intense than the target, or where the local signal-to-noise (S/N) ratio is below some predetermined criterion value. The logic for the parallel between EM and ITFS is based on the assumption that the target information contained in these regions would be "energetically masked" at the ear for a human observer with normal hearing (Brungart *et al.*, 2006, 2009). Previous studies have generally concluded that ITFS is sufficient for achieving this goal. Thus, speech identification thresholds measured for ITFS-processed ("glimpsed") stimuli often are treated as an EM baseline. These baseline thresholds are then compared to thresholds for unprocessed stimuli, and the difference has been taken as an indication of the amount of IM that is present (e.g., Brungart *et al.*, 2006, 2009; Kidd *et al.*, 2016; Kidd *et al.*, 2019; Buss *et al.*, 2017; Rennies *et al.*, 2019). Interpreting this difference as a direct metric of IM, however, is complicated by the fact that, logically, there are certain situations where such a strong conclusion may

[a)]Portions of this work were presented in "A glimpsing model with a variable 'tile' size," 42nd Annual MidWinter Meeting of the Association for Research in Otolaryngology, Baltimore, MD, USA, February 2019.
[b)]Electronic mail: cwconroy@bu.edu
[c)]Also at: Hearing Research Center, Boston University, 44 Cummington Mall, Boston, MA 02115, USA.
[d)]ORCID: 0000-0002-5535-5736.

https://doi.org/10.1121/10.0000893

not be justified. For example, ITFS is typically accomplished using fine T-F resolution, which may yield target "glimpses" (T-F regions of the target + masker mixture that are *not* deemed energetically masked and are included in the glimpsed stimulus) that are too narrow/brief to be resolved individually by the ear in the unprocessed mixture (see Brungart *et al.*, 2006; Kidd *et al.*, 2019). One possible consequence of this could be that estimates of IM obtained via ITFS are inflated because the full effects of EM are not accounted for at baseline (i.e., glimpsed thresholds are effectively "too low" with respect to the EM limits imposed by the ear). More generally, it is possible that thresholds for glimpsed stimuli and, by extension, estimates of IM obtained using these thresholds as a baseline depend directly on the T-F resolution that is employed during ITFS.

ITFS was originally proposed as a tool for separating EM from IM in SOS masking by Brungart *et al.* (2006, 2009). However, Brungart and his colleagues were careful to emphasize that the distinction between EM and IM is by no means straightforward and it is extremely difficult to determine precisely the relative contributions of EM and IM in a situation as complex as SOS masking. This is due, in part, to the fact that performance in the SOS masking task depends on a wide range of complex factors that may not be adequately captured by the simple EM-IM distinction (e.g., linguistic and memory-related processes, expectation and *a priori* knowledge, the integration of speech information over multiple overlapping timescales, phonemic restoration effects, etc.; cf. Warren, 1970; Conway *et al.*, 2001; Mattys *et al.*, 2012; Carlile, 2014; Bronkhorst, 2015; Kidd and Colburn, 2017). Nonetheless, Brungart *et al.* (2006) reasoned that observer success in solving the SOS masking task depends on the completion of a sequence of (at least) two distinct processes: target *detection*, followed by target-masker *segregation* (see also Cooke, 2006; Li and Loizou, 2007; Healy *et al.*, 2014; Healy and Vasko, 2018). By this view, the observer must first detect, via peripheral mechanisms, the T-F regions of the target + masker mixture that are dominated by target energy, and then segregate, via central mechanisms, these target-dominated regions from those that are masker dominated. ITFS ostensibly mimics this sequence of steps "ideally" via computational means such that detection is limited only to the extent that it would be for an ideal observer (i.e., EM limits the input to the observer; cf. Durlach *et al.*, 2003), whereas segregation is not limited at all (i.e., IM is not a factor; Brungart *et al.*, 2006, 2009; see also Kidd *et al.*, 2016; Kidd *et al.*, 2019).

In practice, ITFS is accomplished by passing the target + masker mixture through a bank of n bandpass filters (analogous to auditory frequency filters) and then further subdividing the signal within each band into m-ms time frames (analogous to auditory temporal windows). This process yields a grid of acoustically defined T-F regions ("tiles") that are assumed to correspond, roughly, to the T-F regions that are available internally to the human observer. Next, using *a priori* knowledge about the target and masker waveforms, tiles that are dominated by target energy, or where the local S/N ratio is above some predetermined criterion value, are identified (i.e., *detected*) and then extracted (*segregated*) from among the remaining masker-dominated tiles. The target-dominated T-F tiles (only) are then recombined to form a new signal for use in speech intelligibility experiments. Importantly, masker-dominated tiles are eliminated ("discarded") at this stage of processing, with the assumption being that any target information contained in these tiles would be irrecoverably lost to EM at the ear. It is in this sense that ITFS emulates EM because the thresholds for glimpsed speech are assumed to reflect what the thresholds *would be* in the unprocessed mixture if they were limited *only by EM*. Implicit in this assumption is the further stipulation that in emulating EM, *ITFS also eliminates IM* (Brungart *et al.*, 2006, 2009; Kidd *et al.*, 2016; Kidd *et al.*, 2019) because the same masker-dominated tiles that produce EM—and are discarded during ITFS—also carry information about the speech masker. Thus, removal of masker-dominated tiles during ITFS ostensibly removes the IM that is present in the original unprocessed mixture.

The benefits obtained from ITFS in terms of improving speech intelligibility, therefore, are highly dependent on the extent to which the masker causes IM. For maskers that predominantly produce EM, the benefits of ITFS should, in theory, be negligible. On the other hand, for maskers that predominantly produce IM, the benefits should be large. This general pattern is borne out in the findings of previous studies. For example, Brungart *et al.* (2006) found that the application of ITFS to a low-IM speech-in-noise mixture yielded only a 2–5 dB improvement in thresholds relative to thresholds for unprocessed stimuli, a finding confirmed by Kidd *et al.* (2019) who found a roughly 4 dB benefit of ITFS under similar conditions. On the other hand, these same studies reported that the application of ITFS to a high-IM SOS mixture (a target message masked by one, two, or three colocated same-talker/same-sex speech maskers speaking highly confusable sentences) yielded an improvement in thresholds of as much as 30 dB relative to thresholds for unprocessed stimuli.

A fundamental assumption of the ITFS approach is that the T-F units used in the analysis are a good analog of the processing resolution of the human auditory system and therefore provide a good estimate of EM for a given stimulus configuration. Currently, however, it is not clear how serious a problem there would be for this application of ITFS if there were a significant mismatch in resolution between the acoustic T-F analysis performed on the stimulus and the internal T-F analysis performed by the auditory system. Most previous studies of EM and IM have employed large n and small m, i.e., fine T-F resolution. For example, Brungart *et al.* (2006, 2009), Kidd *et al.* (2016), and Kidd *et al.* (2019) used an n, m pair that we will refer to as the "standard ITFS processing approach": $n = 128$, $m = 20$. It is possible, however, that such fine-grained resolution exceeds the resolution limits of the human auditory system. If this were true, the application of ITFS to a speech mixture would fail to accurately emulate EM by extracting target

glimpses—and including these in the reconstructed stimulus—*that would have been energetically masked at the ear*.

Motivated by this possible mismatch in processing resolution between ITFS and the human auditory system, we sought in this study to determine the influence of the processing parameters *n* and *m* on thresholds for glimpsed speech. Speech identification thresholds were measured for speech and noise maskers after ITFS. ITFS processing resolution was degraded systematically across conditions by varying the parameters *n* and *m* relative to the standard ITFS processing approach. We chose the standard ITFS processing approach as the reference condition because this approach has been used in previous studies of EM and IM using ITFS, and because informal pilot listening suggested a performance plateau as acoustic resolution became more fine grained. It was expected that, consistent with previous studies of ITFS processing resolution not explicitly concerned with the EM-IM distinction (e.g., Li and Loizou, 2008; Montazeri and Assmann, 2018), as the processing resolution decreased, glimpsed thresholds would increase in turn. As the empirical results will show, this expectation was confirmed. The implications of this finding for estimating IM via ITFS are considered, as is the potential role of a mismatch between acoustic and internal T-F resolution. This putative mismatch leads to situations in which "large" (in terms of their T-F extent) tiles created acoustically through ITFS may be further analyzed in the ear or, conversely, fine-grained acoustic resolution provides access to target speech information that would normally be unavailable through auditory processing.

Consideration of the benefits of ITFS for high-EM and high-IM maskers also has important implications for models of masked speech recognition. For example, comparing the intelligibility benefits yielded by ITFS for speech versus noise maskers suggests that when speech is masked by competing speech, IM, rather than EM, dominates performance, a conclusion that is supported by numerous previous studies (e.g., Freyman *et al.*, 1999; Freyman *et al.*, 2004; Brungart, 2001; Brungart *et al.*, 2001; Arbogast *et al.*, 2002; Kidd *et al.*, 2005; Calandruccio *et al.*, 2010; Brouwer *et al.*, 2012; Best *et al.*, 2012; see Kidd and Colburn, 2017, for a review). Current approaches to predicting speech intelligibility under masking, however, fail to take account of these findings or incorporate IM in the predictions. For example, the Articulation Index (AI; French and Steinberg, 1947; Kryter, 1962a,b; see also Egan and Wiener, 1946) and subsequent modifications and extensions of that theory and associated methods [e.g., Speech Intelligibility Index (SII) ANSI, 1997] are based on the computation of S/N ratios in frequency bands originally modeled on the "critical bands" thought to provide the initial frequency analysis of sounds in the auditory periphery (cf. Beranek, 1947). Each band contributes a finite amount of information to overall intelligibility with the contribution depending primarily on the target energy available within each band after taking into account the energy of the masker (i.e., the S/N ratio; although, see the discussion of "remote masking" by Kryter, 1962b). The worst a given band can do is to contribute zero to the AI, meaning that no target information is available. However, the work cited above clearly shows that removing masker-dominated tiles—even when they contain little or no target energy—during ITFS dramatically improves intelligibility contrary to the logic underlying the predictions. That is, the masker-dominated tiles exert a *negative* effect on intelligibility regardless of any target energy present. It is also worth noting that the removal of tiles during ITFS could exert other negative effects on performance for reasons unrelated to EM and IM (at least as defined here). For example, the removal of tiles could cause the disruption of mechanisms producing phonemic restoration (e.g., Warren, 1970; Warren and Obusek, 1971; Bashford *et al.*, 1992). The empirical work described next aims to help resolve these issues and clarify how ITFS can be used to determine the various factors underlying the masking of speech.

II. METHODS

A. Observers

Six observers, including the first author (C.C.), participated (18–29 years of age; mean = 22 years of age). All observers had normal pure-tone air-conduction thresholds at octave frequencies between 250 and 8000 Hz in both ears. Some had participated in previous experiments using ITFS-processed stimuli. All observers received compensation.

B. Stimuli

All stimuli were subjected to ITFS. The generation, processing, and delivery of stimuli were the same as in the "baseline glimpsed" conditions of Kidd *et al.* (speech maskers, Kidd *et al.*, 2016; noise maskers, Kidd *et al.*, 2019), except for the controlled change in *n* and/or *m* in the present study. A close correspondence between the stimuli and methods of Kidd *et al.* (2016), Kidd *et al.* (2019), and the present study was maintained to facilitate a direct comparison of results.

Speech materials were derived from the Boston University Corpus, a laboratory designed corpus of 40 monosyllabic words divided into 5 syntactic categories (name, verb, number, adjective, object), with 8 words in each category (Kidd *et al.*, 2008b). Only female talkers were used. For each trial, two signals were generated prior to ITFS: a target and a masker. The target was a single five-word sentence, always beginning with the cue word "Sue," generated by concatenating one word from each syntactic category in order. A female talker, selected at random (with replacement) from a set of 11, was used to construct the target for each trial. The masker was one of two classes, depending on condition: a speech masker or a noise masker. The speech masker consisted of two additional five-word sentences with the same structure as the target sentence, spoken by two different female talkers (selected from the remaining ten talkers in the set from which the target was drawn). Between the target and masker, all talkers and all words were mutually exclusive (i.e., each sentence used to

https://doi.org/10.1121/10.0000893

construct the speech masker always started with a word other than the cue word "Sue"). The noise masker was speech-spectrum-shaped, speech-envelope-modulated noise. The noise masker was modulated by the single-channel, broadband envelope of a unique, unused speech masker. Envelope extraction was accomplished by passing the unused speech masker through a fourth-order Butterworth low-pass filter with a 300 Hz cutoff. The resulting waveform was multiplied by a speech-shaped Gaussian noise having the same long-term average spectrum as all female talkers in the corpus.

C. Procedures

Observers were tested while seated in a double-walled, sound-treated Industrial Acoustics Company (IAC, North Aurora, IL) booth. The booth had a computer monitor, a keyboard, and a mouse. Digital-to-analog (D/A) conversion was effected on a control computer situated outside of the booth using an RME (Haimhausen, Germany) HDSP 9632 (ASIO) 24-bit sound card, and stimuli were presented binaurally through Sennheiser HD280 Pro headphones (Sennheiser Electronic GmbH and Co. KG, Wedemark, Germany). Observers were told that on each trial their task was to identify the words comprising the five-word target, the target would always be spoken by a female talker, and it would always begin with the cue word "Sue." During the presentation interval of each trial, the computer monitor was blank, and observers were instructed to listen. During the response interval, a 40-word matrix (8×5; exemplar $\times$ syntactic category; cf. Kidd *et al.*, 2008b) of the full corpus appeared on the monitor. Observers were instructed to use the mouse to click each word of the target in order from left to right, one word from each column, always beginning with the cue word "Sue." Observers were forced to select one word from each syntactic category in order with no possibility for corrections. The first word was not scored.

The experiment consisted of 12 total conditions: three *number of bands* conditions (where $n = 8$, 32, or 128), two *duration of time windows* conditions (where $m = 20$ or 80), and two *masker type* conditions (where the masker was speech or noise). For each trial, the root-mean-square (rms) level of the target was fixed at 55 dB sound pressure level (SPL), and the rms level of each individual sentence used to construct the masker was scaled to a predetermined target-to-masker ratio (T/M). Observer performance in each condition was evaluated by collecting data at six different T/Ms, evenly spaced in 5 dB steps. The specific range of T/Ms used to evaluate performance varied with condition and was chosen after pilot listening but fell in the range from -35 to $+5$ dB. Note that, in this study, T/M referred to the level of the target relative to each individual sentence used in the construction of the masker. For example, a T/M of 0 dB described a situation in which each masker sentence was scaled to have the same rms energy as the target or, equivalently, the ratio of the target to the combined masker was roughly -3 dB. S/N ratio was used to describe the latter metric. This convention was adopted to maintain consistency with previous studies of EM and IM using ITFS (e.g., Brungart *et al.*, 2006, 2009; Kidd *et al.*, 2016; Kidd *et al.*, 2019). More specifically, it was adopted to facilitate a direct comparison of our results with those of Kidd *et al.* (2016) and Kidd *et al.* (2019). All signal levels were specified before ITFS.

Prior to the beginning of the initial experimental session, each observer completed a single training block of 12 trials to become familiar with the stimuli and response procedure. All 12 conditions were presented during the training block at a favorable (i.e., > -5 dB) T/M. Immediately following the training block, 864 scored trials were completed. These included 12 trials at each of the 6 T/Ms tested for each of the 12 conditions. The trials were divided into 108 blocks of 8. The condition and T/M tested on each trial were randomized across all trials. Feedback was given on all training and experimental trials. Each observer completed the experiment in two sessions of roughly two hours each.

D. Signal processing

Stimuli were pre-generated and processed digitally at a sampling rate of 44 100 Hz using MATLAB software (MathWorks Inc., Natick, MA). Prior to ITFS, the target and each individual sentence used in the construction of the masker were convolved with a 0° azimuth, laboratory recorded, Knowles Electronic Manikin for Acoustic Research (KEMAR, G.R.A.S. Sound and Vibration, Holte, Denmark) head related transfer function (HRTF). While in all conditions each signal had a nominal spatial position of 0° azimuth, the convolution of each signal with the KEMAR HRTF had the effect of introducing slight differences between the signals at the two ears. Again, this was done primarily to facilitate a direct comparison of our results with those of Kidd *et al.* (2016) and Kidd *et al.* (2019). It is worth noting, however, that convolution with the KEMAR HRTFs, followed by presentation over circumaural headphones may have effectively applied concha/pinna filtering twice and thus introduced frequency specific effects that may have influenced how observers weighted speech information across frequency. As this would not have influenced the T/M, such effects were assumed to be negligible with respect to overall performance. Following convolution, all subsequent processing was carried out on the signal at each ear independently, and the experimental stimulus was resynthesized, post-processing, as a two-channel binaural signal before presentation during the experimental session.

In general, our ITFS processing approach followed closely the approach used in previous studies (e.g., Wang and Brown, 1999; Roman *et al.*, 2003; Wang, 2005; Brungart *et al.*, 2006, 2009; Kidd *et al.*, 2016; Kidd *et al.*, 2016, 2019; Rennies *et al.*, 2019). Specifically, the algorithm used in the present experiment was the same as that used by Brungart *et al.* (2006, 2009), Kidd *et al.* (2016), and Kidd *et al.* (2019) with a few notable exceptions, described below. The reader is referred to those studies for a more thorough discussion of each stage of ITFS processing.

Briefly, in the present study, T-F decomposition was accomplished by passing each signal through a bank of n filters (where $n = 8$, 32, or 128) with overlapping passbands. The signal within each band was then subdivided into m-ms time frames (where $m = 20$ or 80) with 50% overlap. The filters were linearly spaced on the Equivalent Rectangular Bandwidth (ERB) scale between 80 and 8000 Hz, and the bandwidth of each filter was calculated using Eq. (1), an extension of Eq. (3) from Glasberg and Moore (1990, p. 114). For the present experiment, the term n_{ref}/n_{cond} was added:

$$n_{BW}(f) = 24.7(4.37F + 1) \times 1.019 \times (n_{ref}/n_{cond}). \quad (1)$$

In Eq. (1), $n_{BW}(f)$ is the bandwidth of each filter in Hz, F is the filter center frequency in kHz, and 1.019 is the ERB scaling factor. The ERB scaling factor is meant to scale ERB values—as computed by the term $24.7(4.37F + 1)$ (cf. Glasberg and Moore, 1990)—to Gammatone filter bandwidths (cf. Patterson *et al.*, 1992). Because the standard ITFS processing approach was considered to be the reference condition, when calculating filter bandwidths, n_{ref} was always 128, and n_{cond} was the number of frequency bands in the experimental condition under test (i.e., 8, 32, or 128). Note that this added term meant that as n decreased, the bandwidth of the filters increased.

Thus, during ITFS, each signal was decomposed into a two-dimensional matrix of overlapping tiles that varied in their spectro-temporal extent with condition. The within-tile S/N ratio was then calculated for each tile. A local criterion (LC) value of 0 dB (Brungart *et al.*, 2006) was adopted. Tiles in which the S/N ratio $\geq$ LC were considered to be target dominated and were assigned a "1." Tiles in which the S/N ratio $<$ LC were considered to be masker dominated and were assigned a "0." The resultant matrix of 0 s and 1 s formed a binary-mask (Wang, 2005) that was applied to the overall mixture (target + masker). The retained tiles for each ear were resynthesized and stored for presentation as a glimpsed stimulus (for details of the resynthesis process, see Brown and Cooke, 1994; Wang and Brown, 1999; Brungart *et al.*, 2006).

III. RESULTS

Psychometric functions relating percent correct word identification to T/M in dB were obtained for each observer in each condition by fitting the data with a logistic function. Threshold T/M in dB for a given observer/condition was defined as the point at which the fitted function crossed 50% correct. Individual thresholds as well as group means and the associated standard errors of the means for each condition are shown in Table I. Also shown in Table I are the group mean T/Ms at threshold from Kidd *et al.* (speech maskers, Kidd *et al.*, 2016; noise maskers, Kidd *et al.*, 2019) in their "natural baseline" (i.e., unprocessed[1]) conditions (conditions that we did not test in the present study) and in their "baseline glimpsed" conditions (which were equivalent to our $n = 128$, $m = 20$ conditions). Except for differences in n and/or m and observers, the stimuli, methods, and

TABLE I. Individual and group mean thresholds for all conditions tested. n = the number of frequency analysis bands, and m = the duration of the time windows (in ms) used during ITFS. Also shown for comparison are data from Kidd *et al.* (2016) and Kidd *et al.* (2019) in their unprocessed conditions and in conditions that were identical to our $n = 128$, $m = 20$ conditions. See the text for further details.

	Threshold T/M (dB)						
Observer	Unprocessed	*n*8*m*20	*n*8*m*80	*n*32*m*20	*n*32*m*80	*n*128*m*20	*n*128*m*80
Speech masker							
1		−7.7	−6.6	−17.7	−18.4	−31.9	−29.4
2		−16.2	−16.2	−24.4	−25.4	−33.3	−36.8
3		−11.9	−9.5	−18.2	−17.7	−28.4	−29.8
4		−19.7	−11.7	−20.0	−20.5	−32.5	−32.5
5		−8.6	−4.9	−16.5	−17.5	−28.4	−24.2
6		−4.7	−4.1	−12.6	−12.9	−29.4	−24.5
Mean		−11.5	−8.8	−18.2	−18.7	−30.6	−29.5
Standard error		2.3	1.9	1.6	1.7	0.9	2.0
Kidd *et al.* (2016) mean	−0.4					−30.6	
Kidd *et al.* (2016) standard error	1.1					0.8	
Noise masker							
1		−7.5	−7.8	−12.1	−12.7	−18.2	−14.2
2		−13.9	−11.7	−15.8	−11.4	−21.8	−18.2
3		−11.4	−8.5	−11.4	−8.9	−16.7	−13.0
4		−11.5	−9.5	−13.7	−14.0	−20.5	−15.8
5		−5.1	−6.5	−10.7	−5.8	−17.0	−11.7
6		−3.3	−4.2	−9.4	−3.8	−13.7	−12.2
Mean		−8.8	−8.0	−12.2	−9.4	−18.0	−14.2
Standard error		1.7	1.0	0.9	1.6	1.2	1.0
Kidd *et al.* (2019) mean	−13.5					−17.8	
Kidd *et al.* (2019) standard error	1.1					0.7	

https://doi.org/10.1121/10.0000893

procedures used by Kidd *et al.* (2016) and Kidd *et al.* (2019) were identical to those used in the present study, and thus a direct comparison of our results with theirs is appropriate. These data are discussed in Sec. IV C.

Group mean T/Ms at threshold and the associated standard errors of the means from Table I are plotted in Fig. 1 for speech maskers (left) and noise maskers (right). Visual inspection of Fig. 1 reveals three clear trends. First, thresholds for noise maskers were, on average, higher than those for speech maskers, in general agreement with previous studies comparing thresholds for speech and noise maskers after ITFS (Brungart *et al.*, 2009; Kidd *et al.*, 2019). Second, as n decreased, thresholds increased, regardless of the value of m or masker type. This effect was greater for speech maskers than it was for noise maskers, in general agreement with previous studies that have examined the role of frequency resolution during ITFS not explicitly concerned with the EM-IM distinction (e.g., Li and Loizou, 2008; Montazeri and Assmann, 2018). Third, increasing m at a given n had a relatively minor—yet, consistently negative—influence on thresholds compared to the more pronounced effect of reduced frequency resolution.

The individual threshold data shown in Table I and Fig. 1 were subjected to a three-factor [(*number of bands*) × (*duration of time windows*) × (*masker type*)], within-subjects analysis of variance. Consistent with the trends noted above, all three main effects were significant: There was a significant main effect of the *number of bands* [$F(2,10) = 259.69$, $p < 0.0001$], *duration of time windows* [$F(1,5) = 40.48$, $p < 0.01$], and *masker type* [$F(1,5) = 227.93$, $p < 0.001$]. The two-way interaction term [(*number of bands*) × (*masker type*)] also was significant [$F(2,10) = 84.87$, $p < 0.0001$], as was the three-way interaction between all factors [$F(2,10) = 4.70$, $p < 0.05$]. That is, changes to n and/or m influenced observer performance differently depending on masker type. Therefore, the data for speech and noise maskers were analyzed separately. For speech maskers, only the main effect of the *number of bands* was significant [$F(2,10) = 239.81$, $p < 0.0001$]; for noise maskers, both main effects were significant [*number of bands*, $F(2,10) = 87.22$, $p < 0.0001$; *duration of time windows*, $F(1,5) = 46.91$, $p < 0.01$]. Effect size calculations (generalized eta-square, η_G^2; Olejnik and Algina, 2003) indicated that the relative effect of n was greater for speech maskers than for noise maskers ($\eta_G^2 = 0.81$ versus $\eta_G^2 = 0.56$ for speech versus noise maskers, respectively), and for noise maskers, the effect of m was small compared to the effect of n ($\eta_G^2 = 0.15$ versus $\eta_G^2 = 0.56$ for m versus n, respectively).

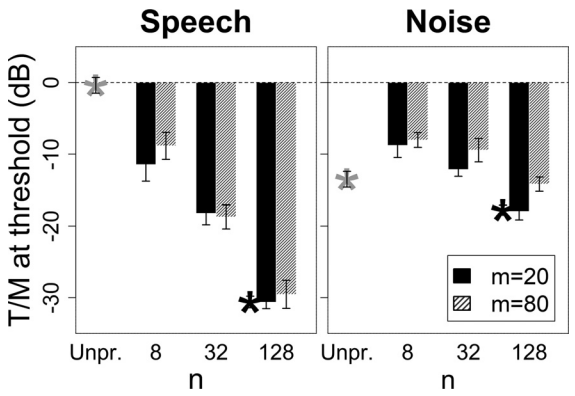

FIG. 1. Group mean T/Ms at threshold in dB. Error bars are standard errors of the means. (Left) speech maskers, (right) noise maskers. The ordinate gives the T/M at threshold in dB. The abscissa gives the value of n. Black bars represent conditions in which $m = 20$. Grey bars represent conditions in which $m = 80$. Data from Kidd *et al.* (speech maskers, Kidd *et al.*, 2016: noise maskers, Kidd *et al.*, 2019) are shown as asterisks for comparison. Black asterisks mark thresholds obtained in conditions identical to our $n = 128$, $m = 20$ conditions. Grey asterisks mark thresholds for unprocessed (yet otherwise identical) stimuli, which we did not test.

Because we failed to observe a significant main effect of m for speech maskers, thresholds at each n were collapsed across m, and a series of paired-sample t-tests were conducted to explore the effect of n in more detail. All comparisons between and among levels of n for speech maskers were highly significant ($p < 0.0001$; Holm-Bonferroni correction). That is, observer performance deteriorated significantly with each subsequent reduction in frequency resolution relative to the standard ITFS processing approach, and the specific value of m that was chosen did little either to contribute to or counteract this effect. For noise maskers, the same series of comparisons between and among levels of n yielded similar results ($p < 0.01$ for all comparisons; Holm-Bonferroni correction, as above).

IV. DISCUSSION

A. How does ITFS processing resolution influence thresholds for glimpsed speech?

The two findings that most directly answered this question were the following: First, reduced frequency resolution (lower n) relative to the standard ITFS processing approach had a significant negative influence on thresholds for both speech and noise maskers, although the magnitude of the effect was greater for speech maskers (thresholds increased by roughly 20 dB for $n = 8$ versus $n = 128$, averaged across m) than it was for noise maskers (a roughly 8 dB increase for the same comparison). Second, reduced temporal resolution relative to the standard ITFS processing approach (an increase in m from 20 to 80 ms) had only a slight negative influence on thresholds, which was greater for noise maskers (statistically significant) than it was for speech maskers (not statistically significant). Taken together, these two broad findings suggest that n—as opposed to m—is the dominant ITFS processing parameter in determining thresholds for glimpsed speech. We did not, however, test values of $m > 80$, and it is possible that if we had we would have seen a larger effect of this parameter. Similarly, the standard ITFS processing approach provided the upper bound on T-F resolution in the present study, and for both speech and noise maskers, thresholds for stimuli generated using this approach were lower than thresholds for stimuli generated using any other n, m pair. It is possible that processing resolution more fine-grained than the standard ITFS processing

approach would have yielded even lower thresholds still. While informal pilot listening in our laboratory suggests that this is not the case, further empirical work is needed to determine which n, m pair outside of the range of values tested here yields best performance. It is worth noting, however, that resolution which is more fine-grained than the standard ITFS processing approach likely would exceed the resolution limits of the human auditory system, and thus would be of little value when ITFS is used to separate EM from IM in speech masking experiments.

The results reported here are generally consistent with previous studies of frequency resolution during ITFS. For example, Li and Loizou (2008) measured percent correct word identification for both ITFS-processed and unprocessed stimuli. Speech masked by either multitalker babble or a steady-state speech-shaped noise was synthesized using a sinewave vocoder with n channels ($n = 6$, 12, 16, 24, or 32), and ITFS was applied to those channels using 4-ms time frames ($m = 4$). As in the present study, they found that glimpsed stimuli remained highly intelligible for all $n \geq 12$. At low S/N ratios, when $n = 24$ or 32, the intelligibility benefit of ITFS relative to unprocessed (vocoded) stimuli was large, particularly when the masker was multitalker babble (roughly 60 percentage points). However, when $n < 12$, a much smaller benefit was found. Montazeri and Assmann (2018) also examined the role of frequency resolution during ITFS using vocoded stimuli and a processing scheme similar to that used by Li and Loizou (2008). They tested $n = 6$ or 12. They found that speech recognition suffered significantly as n decreased from 12 to 6. This effect was particularly pronounced for speech maskers (a single competing talker) versus noise maskers (steady-state speech-shaped noise). Thus, the common theme that emerges from these two studies and ours is that large values of n typically yield better performance than small values of n, and the effects of decreasing n relative to a large-n reference are typically greater for speech maskers than they are for noise maskers.

To our knowledge, no previous studies have directly examined the role of temporal resolution during ITFS in a manner consistent with the present study, although Li and Loizou (2007) collected relevant data in this area. Our finding of a slightly negative influence of reduced temporal resolution in noise, however, is broadly consistent with previous studies of "glimpsing" and interrupted speech more generally (i.e., studies not limited to ITFS), in which it is found that shorter, more frequent glimpses typically support better intelligibility than longer, less frequent glimpses (e.g., Miller and Licklider, 1950; Huggins, 1975; Li and Loizou, 2007; Wang and Humes, 2010; Gibbs and Fogerty, 2018).

B. Possible explanations for the increase in thresholds with the reduced ITFS processing resolution

Before addressing this issue, three terms are defined here. They are "target-dominated tiles-within-tiles" (TD-TWTs), "masker-dominated tiles-within-tiles" (MD-TWTs), and "lost tiles" (LTs). Figure 2 illustrates the stimulus

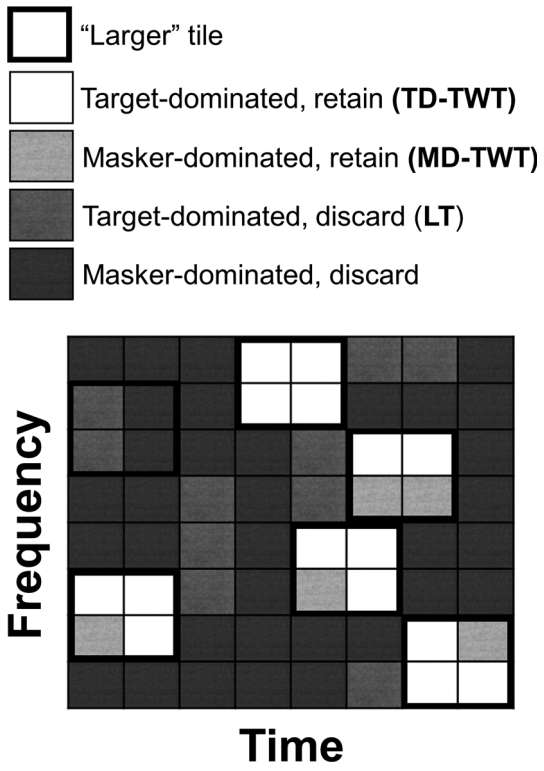

FIG. 2. Schematic showing possible categories of subtiles within ITFS-processed stimuli. See the text for further details.

categories corresponding to these three terms in schematic form. Each of these terms refers to a hypothetical class of subtile that was not directly manipulated in the present study and pertains only to conditions in which $n < 128$ and/or $m = 80$. TD-TWTs refers to "smaller" tiles (defined in terms of their T-F extent by the standard ITFS processing approach) that were target dominated and fell within a "larger" tile (defined by any n, m pair other than the standard ITFS processing approach) that was target dominated overall. MD-TWTs refers to the complementary phenomenon: smaller tiles that were masker dominated, yet, fell within a larger tile that was target dominated overall. Both TD-TWTs and MD-TWTs were retained after ITFS and included in the reconstructed glimpsed stimulus when $n < 128$ and/or $m = 80$ regardless of their local (subtile) S/N. Conversely, LTs were discarded during ITFS. The term LTs refers to smaller tiles that were target dominated yet fell within a larger tile that was masker dominated overall. Thus, LTs represent tiles that were removed during ITFS when $n < 128$ and/or $m = 80$, yet, that *would have been retained* had the acoustic mixture been analyzed using the standard ITFS processing approach.

1. Explanation in terms of an increase in IM

One possible explanation for the increase in thresholds with decreasing n (and for noise, increasing m) is that there was an effective reintroduction of IM via the MD-TWTs. That is, it is possible that when the ITFS analysis tile was large (when $n < 128$ and/or $m = 80$), observers were able to "re-glimpse" the stimulus within each tile internally to

https://doi.org/10.1121/10.0000893

further separate TD-TWTs from MD-TWTs, thereby isolating potentially confusable bits of the masker that could cause IM. Under the framework of ITFS, this would amount to a failure to eliminate IM via ITFS. Note that this putative re-glimpsing process assumes that internal resolution was more fine-grained than acoustic resolution in these conditions. This interpretation did not receive strong support. While the magnitude of the increase in thresholds with decreasing n was greater for speech maskers than it was for noise maskers (cf. Fig. 1), consistent with the prediction of a reintroduction of IM via the MD-TWTs (i.e., MD-TWTs from a noise masker presumably would cause little IM because they are perceptually dissimilar to the TD-TWTs, e.g., Kidd *et al.*, 2005), this result is not compelling, in part, because the proportion of all single word identification errors that were explicit confusions between the target word and a concurrent masker word never rose significantly above chance in any speech masker condition tested ($p > 0.05$; see Brungart, 2001; Brungart *et al.*, 2006; Ihlefeld and Shinn-Cunningham, 2008; Iyer *et al.*, 2010; Kidd *et al.*, 2016, for the logic behind this analysis). Based on this result, our conclusion was that the increase in thresholds with reduced ITFS processing resolution was not attributable to a reintroduction of IM. It should be noted, however, that explicit masker confusions are neither necessary nor sufficient to indicate the presence of IM in a speech mixture (e.g., Kidd *et al.*, 2016), and therefore we cannot rule out the possibility that a reintroduction of IM via MD-TWTs was a factor in the present study.

2. Explanation in terms of a loss of target information

Another, more straightforward, explanation is that as the ITFS processing resolution was reduced relative to the standard ITFS processing approach, there was a decrease in the amount of target energy that was retained in the glimpsed stimulus following ITFS, yielding an increase in thresholds in turn. Under the framework of ITFS, this would amount to an increase in EM. Acoustic analyses support this interpretation. Figure 3 shows how two complementary acoustic metrics varied as a function of T/M in each condition tested. The top row shows the proportion of target energy *retained* in the glimpsed stimulus after ITFS (the proportion of target energy in the glimpsed stimulus relative to the unprocessed target; cf. Brungart *et al.*, 2009; Kidd *et al.*, 2016; Kidd *et al.*, 2019), whereas the bottom row shows the proportion of target energy *discarded* with the LTs (the proportion of target energy discarded with the LTs relative to the unprocessed target[2]). The lines are functions fit to estimates of these two metrics,[3] calculated using 50 randomly generated exemplars at 9 different T/Ms (−35–5 dB in 5 dB steps). The symbols on each line are located at the behavioral group mean T/M at threshold for the condition corresponding to that line. In the top row, the point at which each line crosses the thick, dashed horizontal line represents "threshold effective T/M" (Brungart *et al.*, 2009, p. 4015), defined in this study as the T/M at which the same proportion of target energy was retained in the glimpsed stimulus in each condition as was retained at threshold in the reference condition.[4]

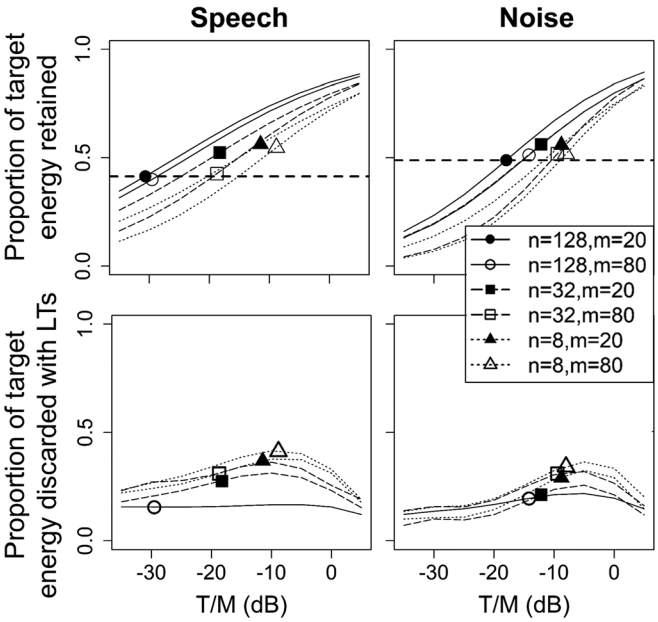

FIG. 3. Estimated proportion of target energy retained (top row) and target energy discarded with the LTs (bottom row) after ITFS as a function of T/M in each condition tested. The ordinate gives the ratio of energy retained/discarded to the energy of the unprocessed target. Lines show best fitting functions to the data, fit separately for each condition. The symbols on each line are located at the behavioral group mean T/M at threshold for the condition corresponding to that line (cf. Table I). The thick, dashed horizontal line in the top two panels marks the proportion of target energy that was retained at the group mean behavioral threshold in the reference condition.

Consider first the proportion of target energy retained functions shown in the top row of Fig. 3. For both speech and noise maskers, the standard ITFS processing approach yielded the greatest proportion of target energy retained across T/Ms. It also yielded the lowest behavioral thresholds. Relative to this energetic baseline, decreasing n—and increasing m at a given n—resulted in a decrease in the amount of target energy retained, much like how the same manipulations often resulted in an increase in behavioral thresholds. For example, for speech maskers, threshold effective T/M increased from −29.6 dB for $n = 128$ (solid lines) to −22.0 dB for $n = 32$ (dashed lines) and to −17.7 dB for $n = 8$ (dotted lines) (averaged across m), whereas behavioral thresholds increased from −30.0 dB to −18.5 dB and to -10.2 dB for the same manipulations of n (also averaged across m). In other words, as n decreased relative to the standard ITFS processing approach, increasing amounts of target energy were discarded during processing, and this likely contributed to the increase in thresholds evident in the behavioral data. Inspection of these functions makes clear, however, that differences in energy retained across conditions did not fully explain differences in thresholds. For example, for $n < 128$ speech masker conditions, thresholds were generally poorer than would be predicted by the "constancy hypothesis" (e.g., Kidd *et al.*, 2019, p. 454), which predicts that threshold-level speech identification

performance requires a fixed proportion of target energy retained in the glimpsed stimulus (i.e., thresholds across conditions would fall along the threshold effective T/M line). By contrast, thresholds for noise maskers were much closer to what would be predicted by the constancy hypothesis.

There are many ways in which the speech and noise maskers differed that may have influenced the distribution of glimpses across the T-F plane (e.g., the noise maskers were primarily temporally, rather than spectro-temporally, modulated), and this too likely played a role in performance (e.g., Buss *et al.*, 2009; Wang and Humes, 2010; Shafiro *et al.*, 2011; Gibbs and Fogerty, 2018; Kidd *et al.*, 2019). For example, with respect to the thresholds for speech maskers, it is possible that decreasing n negatively influenced thresholds by degrading temporal aspects of the stimulus that are crucial for intelligibility (e.g., the variation in temporal envelopes within frequency channels; Shannon *et al.*, 1995). More generally, it is possible that as n decreased, increasingly large "chunks" of speech information were discarded during ITFS, making it more difficult for observers to perceptually string together individual glimpses across time.

3. Importance of LTs

The discussion above indicates that differences in energy retained across conditions were an important factor in performance. Under the framework of ITFS, differences in energy retained suggest differences in EM. Viewed in this way, LTs are important because they contain target information that is available when the ITFS processing resolution is sufficiently fine grained to recover that information from the target + masker mixture but that is lost when resolution is more coarse. As such, the LTs functions, shown in the bottom row of Fig. 3, indicate the degree to which the standard ITFS processing approach *underestimates EM* for a given stimulus configuration, assuming a perfect match between acoustic and internal resolution for each other n, m pair. By way of example, consider the function for the $n = 32$, $m = 20$ speech masker condition (the dashed line with a black square in the bottom left panel of Fig. 3) and its associated behavioral threshold. If we assume a hypothetical observer whose internal T-F resolution perfectly matches the acoustic T-F resolution that was used in this condition, this function/threshold suggests that stimuli generated with the standard ITFS processing approach contained, on average, roughly 30% *more* target energy than would be available to this observer internally in an unprocessed SOS mixture. That is, the standard ITFS processing approach would underestimate EM by roughly 30% for this observer, thereby potentially inflating estimates of IM in turn. To the extent that the standard ITFS processing approach exceeds the resolution limits of the human auditory system, it is possible, then, that thresholds for stimuli generated with this approach reflect the *use of information contained in the LTs*, and therefore these tiles may directly influence estimates of IM.

It is important to keep in mind that LTs are a hypothetical variable, and we do not, at present, have corresponding empirical work to support our speculation about the possible role/importance of the energy contained in the LTs. However, Fig. 4 shows a significant, positive correlation between the proportion of target energy discarded with the LTs and observer thresholds across conditions (speech maskers, $R^2 = 0.83$, $p < 0.01$; noise maskers, $R^2 = 0.71$, $p < 0.01$), implying that, when LTs were retained (i.e., in the reference condition), the information contained in those tiles contributed to intelligibility. This has implications for characterizing susceptibility to IM in observers with sensorineural hearing loss, where a mismatch between acoustic and internal resolution would likely be an even greater issue due to the poorer frequency selectivity typically associated with this observer group (cf. Kidd *et al.*, 2019). As a rough estimate of the possible magnitude of this effect, a reduction by one-half of the number of available frequency bands (e.g., from 32 to 16 bands; cf. Fig. 1 and Table I, for $m = 20$ or $m = 80$) would elevate glimpsed thresholds by as much as 4–5 dB, leading to an overestimate of IM by that amount.

C. Implications for estimates of IM

It may be concluded from scrutiny of the threshold data that estimates of IM obtained via ITFS strongly depend on the ITFS processing resolution. That is, under the framework of ITFS, when thresholds for glimpsed stimuli vary—as they did in the present study with n (and in some cases m)—so too do estimates of IM. Although we did not measure thresholds for unprocessed stimuli and therefore cannot estimate IM directly, we can get a sense of how the ITFS processing resolution influences estimates of IM by calculating IM at the group level using thresholds for unprocessed stimuli obtained by Kidd *et al.* (2016) and Kidd *et al.* (2019)

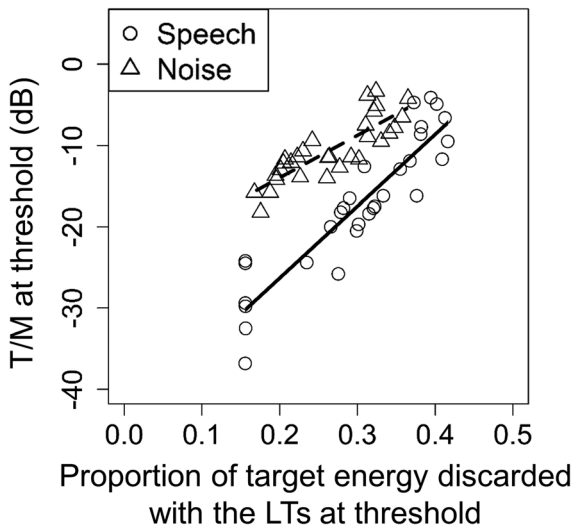

FIG. 4. Behavioral T/Ms at threshold for each observer in each condition (other than the reference condition; cf. Table I) plotted against the proportion of target energy that was discarded with the LTs at threshold in the corresponding condition. Speech masker conditions are shown as circles and noise masker conditions are shown as triangles. The proportion of target energy discarded values was obtained on a per-observer basis using the functions shown in the bottom row of Fig. 3. Lines are linear least-squares fits to the data for speech maskers (solid line) and noise maskers (dashed line).

https://doi.org/10.1121/10.0000893

using similar methods as a reference. For a given masker type, this amounts to taking the difference between the group mean T/M at threshold for unprocessed stimuli from Kidd *et al.* (2016) and Kidd *et al.* (2019) (grey asterisks in Fig. 1) and the group mean threshold found in each condition tested here. For speech maskers, this approach yields 29.6 dB, 18.1 dB, and 9.8 dB of IM for $n = 128$, $n = 32$, and $n = 8$, respectively (averaged across m, in keeping with the behavioral results). Clearly, estimates of IM are highly dependent on ITFS processing resolution.

The fact that estimates of IM depend on ITFS processing resolution suggests that accurately estimating IM via ITFS depends on the extent to which acoustic and internal T-F resolutions match. One clue as to how such a match might be achieved comes from the noise masker data. For noise maskers, we found that the standard ITFS processing approach yielded glimpsed thresholds that were, on average, 4 dB lower than thresholds for identical unprocessed stimuli obtained by Kidd *et al.* (2019), suggesting that, had we measured thresholds for unprocessed stimuli here, we would have found roughly 4 dB of IM in this condition. Assuming a perfect match between acoustic and internal resolution, however, *no IM should be present at all* for a masker that (1) produced only EM and (2) was independent across tiles. In fact, the only outcome of ITFS would be a *decrease* in intelligibility due to a loss of target information, even if that information contributed minimally to overall intelligibility. Thus, positive IM for noise maskers suggests that either these stimuli violated one or both of these assumptions, or there was a mismatch between acoustic and internal resolution when $n = 128$, $m = 20$. With respect to (1), there is certainly evidence that Gaussian noise maskers, especially when modulated by speech envelopes, exert a masking effect not wholly consistent with the narrow definition of EM adopted here (e.g., Stone *et al.*, 2012; Stone and Moore, 2014; Schubotz *et al.*, 2016; see Culling and Stone, 2017, for a review), and thus the application of ITFS may reduce masking effects not typically associated with IM. With respect to (2), it is possible that adjacent tiles interact such that masker energy from one tile "spills over" into another, violating the criterion of independence. By this view, one way ITFS may improve performance in a speech-in-noise mixture is by reducing forward and/or backward masking of target glimpses that are temporally adjacent to masker-dominated tiles (e.g., Brungart *et al.*, 2006).

Irrespective of the extent to which one or both of these assumptions was violated, it is possible (indeed likely) that the standard ITFS processing approach exceeds the resolution limits of the human auditory system. Whereas positive IM for noise maskers suggests acoustic resolution that is too fine grained, *negative* IM suggests the opposite. For example, in some of the noise masker conditions tested ($n = 32$, $m = 28$; $n = 8$, $m = 20$; and $n = 8$, $m = 80$), thresholds were poorer than the thresholds for unprocessed stimuli obtained by Kidd *et al.* (2019), suggesting negative IM in these conditions. Negative IM for noise is consistent with the expectation of an EM-dominated condition in which acoustic T-F resolution is poorer than internal resolution, resulting in a loss of target energy during ITFS. This suggests that the point at which ITFS yields *no IM* for noise, either positive or negative, can be taken as a crude estimate of internal T-F resolution, or at least where the best estimates of EM are obtained within the constraints of the method. Thresholds in the $n = 32$, $m = 20$ and $n = 128$, $m = 80$ noise masker conditions were both very close (roughly 1 dB) to the Kidd *et al.* (2019) thresholds for unprocessed noise-masked stimuli (cf. Table I and Fig. 1), which suggests that we would have obtained roughly 0 dB of IM in these conditions had we measured the necessary thresholds. This finding, taken together with the fact that roughly 30 bandpass filters are often used to model auditory frequency selectivity across the speech frequency range (e.g., Zwicker, 1961; Zwicker and Scharf, 1965; Moore and Glasberg, 1983; Glasberg and Moore, 1990), suggests that $n = 32$, $m = 20$ may be the most reasonable choice of parameters for researchers seeking to estimate IM via ITFS. One caveat here is that, in the present study, n and bandwidth covaried. That is, as n decreased, bandwidth increased [cf. Eq. (1)]. When $n = 128$, the filters were typical Gammatone filters; when $n = 32$, the filters had bandwidths that were wider than typical Gammatone filters. Thus, the frequency selectivity provided by $n = 32$, $m = 20$ in the present study was not wholly analogous to common models of the ear.

Another factor to consider is that removing the masker dominated tiles in the noise masker conditions also reduced phonemic restoration effects. That is, when the noise was present in the unprocessed stimulus used by Kidd *et al.* (2019), speech intelligibility could have been enhanced because the noise contributed to phoneme recognition via the well-known mechanism of perceptual restoration (e.g., Warren, 1970). The magnitude of this effect is difficult to ascertain, however, in the current conditions because we used ITFS to remove the preponderance of the noise masker in every case. For our data, at least, speech intelligibility enhancement due to perceptual restoration was likely a factor only in the comparisons to the full unprocessed noise masker case used in Kidd *et al.* (2019). However, a study specifically intended to examine this potential effect under the full set of (un)processed conditions appears to be needed to adequately address this issue.

D. Implications for speech in "noise" prediction

IM poses a particularly difficult challenge for the prediction of speech intelligibility under masking. The classic approaches based on AI theory (cf. French and Steinberg, 1947; Kryter, 1962a,b), such as the SII (ANSI, 1997) or modified versions of the SII seeking to account for modulation (e.g., speech transmission index, STI; Steeneken and Houtgast, 1980), scale the contributions of the individual units (frequency bands, typically) between zero and some proportion so that the overall intelligibility maximum is one. This means that each unit contributes zero or more toward overall performance. IM inherently means that some units

make a *negative* contribution to intelligibility, and this becomes particularly obvious when one considers the fact that the removal of masker-dominated tiles during ITFS often improves intelligibility under high-IM conditions, contrary to the predictions of AI-based models. Thus, the problem for such models is how to take IM into account when making predictions. Currently, the most feasible approach would appear to involve incorporating negative weights for certain T-F units in arriving at the overall prediction. This work, which relies on estimates of IM obtained using ITFS, is ongoing.

Finally, it should be reiterated that the EM-IM distinction is complicated, and there are inherent limitations to an ITFS-based analysis of SOS masking performance. For example, the framework of ITFS assumes that speech intelligibility under masking is reasonably well described by a highly simplified two-stage sequence of steps consisting of (1) detection of target-dominated T-F units in a mixture followed by (2) segregation of these units from the interfering background (cf. the Introduction and Brungart *et al.*, 2006). Thus, the informational substrate of intelligibility is assumed to be the individual target-dominated units. However, speech information (e.g., at the phonemic, word, and sentence level) is distributed widely across both time and frequency, and the extent to which the resolution of ITFS matches the "resolution of speech intelligibility" (i.e., the T-F windows over which human observers integrate speech information) may be an important factor to consider, one that is completely ignored here. For example, Saberi and Perrott (1999) showed that intelligibility for sentences remained near-ceiling when the sentences were time reversed in 50-ms segments, suggesting a dissociation between speech information at the level of a T-F unit and some larger temporal integration window. The fact that intelligibility is similarly resilient to degradations in the frequency domain (e.g., noise vocoding; Shannon *et al.*, 1995) further supports such a dissociation.

V. SUMMARY AND CONCLUSIONS

Speech identification thresholds were measured for both speech and noise maskers after ITFS. The ITFS processing resolution was degraded systematically across conditions by varying the parameters n (number of frequency analysis bands) and m (duration of the time windows in ms) relative to "the standard ITFS processing approach" ($n = 128$, $m = 20$). The findings can be summarized as follows:

(1) The number of frequency analysis bands applied during ITFS had a highly significant negative influence on thresholds for glimpsed speech. For all $n < 128$, thresholds were poorer than thresholds for stimuli generated using the standard ITFS processing approach, regardless of m or masker type. The magnitude of the effect was substantial, resulting in an increase in thresholds of roughly 20 dB in the most extreme case ($n = 8$ versus $n = 128$ for speech maskers, averaged across m). A loss of viable target energy appeared to be the primary factor responsible for this effect.

(2) The duration of the time windows applied during ITFS had a relatively minor influence on thresholds. An increase in m from 20 to 80 ms at a given n yielded significantly higher thresholds for noise maskers only. However, the trends in the data, though slight, revealed a systematic negative influence of this parameter on thresholds for both speech and noise maskers. As with changes in n, a loss of target energy likely was a factor in this performance decrement.

(3) The extent to which ITFS processing resolution matches the resolution limits of the human auditory system may, in theory, affect thresholds for glimpsed stimuli and, in turn, estimates of IM obtained via ITFS. This could be due to a variety of factors, including usable target information that is lost when larger T-F tiles are discarded during ITFS and usable target energy that is retained when it normally would be masked.

(4) In support of (3), we observed a slight decrease in thresholds for stimuli generated using the standard ITFS processing approach relative to thresholds for identical unprocessed stimuli measured in a previous study (Kidd *et al.*, 2019) *even in low-IM conditions*, suggesting that the standard ITFS processing approach, with its fine T-F resolution, may not be a sufficient control for EM in speech masking experiments. That is, the standard ITFS processing approach may yield spurious estimates of IM by failing to fully capture the effects of EM at baseline. Our results suggest that $n = 32$, $m = 20$ may yield more accurate estimates of IM due to a potentially closer correspondence with auditory frequency selectivity.

(5) The presence and degree of IM—especially in SOS masking conditions—suggests that models predicting masked speech intelligibility should be modified to include negative weights to account for conditions high in IM. Estimates of IM obtained via ITFS likely will make an important contribution to this project.

ACKNOWLEDGMENTS

This work was supported by NIH-NIDCD (the National Institutes of Health-the National Institute on Deafness and Other Communication Disorders) Grant Nos. R01DC004545 and T32DC013017. The authors wish to thank Christine R. Mason for helpful discussions and technical assistance during the preparation of this manuscript.

[1]Unprocessed in the sense that the LC value used during ITFS was $-\infty$, resulting in an experimental stimulus that was perceptually similar to the unprocessed mixture signal, yet retained any acoustic distortions or artifacts that may have occurred during processing (cf. Brungart *et al.*, 2006).

[2]Rough estimates of the proportion of target energy discarded with the LTs were obtained by first analyzing the overall mixture using a given n,m combination, identifying the discarded tiles, and then "re-glimpsing" these discarded tiles using the standard ITFS processing approach by further subdividing the region of the discarded tile into the appropriate number of subbands (e.g., 4 for $n = 32$ and 16 for $n = 8$) and time epochs (e.g., 4 for $m = 80$). Due to the specifics of the ITFS processing approach

used here (e.g., 50% temporal overlap of contiguous tiles within a frequency band), this is not a perfect approach and thus these estimates should, again, be considered *rough* estimates. Nonetheless, these functions capture the essence of the pattern of energy discarded with the LTs across conditions.

[3]The proportion of target energy retained data was fit separately in each condition with a beta regression model (Ferrari and Cribari-Neto, 2004). The proportion of target energy discarded with the LTs data was fit separately in each condition with a least-squares polynomial.

[4]This is a slightly different definition of threshold effective T/M than used by Brungart *et al*. (2009), who defined threshold effective T/M as a fixed proportion (0.20) of target energy retained.

ANSI (**1997**). ANSI S3.5-1997, *American National Standard: Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, Melville, NY).

Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**(5), 2086–2098.

Bashford, J. A., Riener, K. R., and Warren, R. M. (**1992**). "Increasing the intelligibility of speech through multiple phonemic restorations," Percept. Psychophys. **51**(3), 211–217.

Beranek, L. L. (**1947**). "The design of speech communication systems," Proc. IRE **35**(9), 880–890.

Best, V., Marrone, N., Mason, C. R., and Kidd, G., Jr. (**2012**). "The influence of non-spatial factors on measures of spatial release from masking," J. Acoust. Soc. Am. **131**(4), 3103–3110.

Bronkhorst, A. W. (**2015**). "The cocktail-party problem revisited: Early processing and selection of multi-talker speech," Atten. Percept. Psychophys. **77**(5), 1465–1487.

Brouwer, S., Van Engen, K. J., Calandruccio, L., and Bradlow, A. R. (**2012**). "Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content," J. Acoust. Soc. Am. **131**(2), 1449–1464.

Brown, G. J., and Cooke, M. (**1994**). "Computational auditory scene analysis," Comput. Speech Lang. **8**(4), 297–336.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**(3), 1101–1109.

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**(6), 4007–4018.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2009**). "Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers," J. Acoust. Soc. Am. **125**(6), 4006–4022.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**(5), 2527–2538.

Buss, E., Leibold, L. J., Porter, H. L., and Grose, J. H. (**2017**). "Speech recognition in one- and two-talker maskers in school-age children and adults: Development of perceptual masking and glimpsing," J. Acoust. Soc. Am. **141**(4), 2650–2660.

Buss, E., Whittle, L. N., Grose, J. H., and Hall, J. W. III (**2009**). "Masking release for words in amplitude-modulated noise as a function of modulation rate and task," J. Acoust. Soc. Am. **126**(1), 269–280.

Calandruccio, L., Dhar, S., and Bradlow, A. R. (**2010**). "Speech-on-speech masking with variable access to the linguistic content of the masker speech," J. Acoust. Soc. Am. **128**(2), 860–869.

Carlile, S. (**2014**). "Active listening: Speech intelligibility in noisy environments," Acoust. Aust. **42**(2), 90–96.

Conway, A. R., Cowan, N., and Bunting, M. F. (**2001**). "The cocktail party phenomenon revisited: The importance of working memory capacity," Psychon. Bull. Rev. **8**(2), 331–335.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**(3), 1562–1573.

Culling, J. F., and Stone, M. A. (**2017**). "Energetic masking and masking release," in *The Auditory System at the Cocktail Party, Springer Handbook of Auditory Research 60*, edited by J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay (Springer, New York, pp. 41–73).

Durlach, N. I., Mason, C. R., Kidd, G., Jr., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking (L)," J. Acoust. Soc. Am. **113**(6), 2984–2987.

Egan, J. P., and Wiener, F. M. (**1946**). "On the intelligibility of bands of speech in noise," J. Acoust. Soc. Am. **18**(2), 435–441.

Ferrari, S., and Cribari-Neto, F. (**2004**). "Beta regression for modelling rates and proportions," J. Appl. Stat. **31**(7), 799–815.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**(1), 90–119.

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**(5), 2246–2256.

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**(6), 3578–3588.

Gibbs, B. E., and Fogerty, D. (**2018**). "Explaining intelligibility in speech-modulated maskers using acoustic glimpse analysis," J. Acoust. Soc. Am. **143**(6), EL449–EL455.

Glasberg, B. R., and Moore, B. C. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**(1-2), 103–138.

Healy, E. W., and Vasko, J. L. (**2018**). "An ideal quantized mask to increase intelligibility and quality of speech in noise," J. Acoust. Soc. Am. **144**(3), 1392–1405.

Healy, E. W., Youngdahl, C. L., and Apoux, F. (**2014**). "Evidence for independent time-unit processing of speech using noise promoting or suppressing masking release," J. Acoust. Soc. Am. **135**(2), 581–584.

Huggins, A. W. F. (**1975**). "Temporally segmented speech," Percept. Psychophys. **18**(2), 149–157.

Ihlefeld, A., and Shinn-Cunningham, B. (**2008**). "Spatial release from energetic and informational masking in a selective speech identification task," J. Acoust. Soc. Am. **123**(6), 4369–4379.

Iyer, N., Brungart, D. S., and Simpson, B. D. (**2010**). "Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task," J. Acoust. Soc. Am. **128**(5), 2998–3010.

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (**2008a**). "Informational masking," in *Auditory Perception of Sound Sources*, *Springer Handbook of Auditory Research*, 29, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer, New York), pp. 143–189.

Kidd, G., Jr., Best, V., and Mason, C. R. (**2008b**). "Listening to every other word: Examining the strength of linkage variables in forming streams of speech," J. Acoust. Soc. Am. **124**(6), 3793–3802.

Kidd, G., Jr., and Colburn, H. S. (**2017**). "Informational masking in speech recognition," in *The Auditory System at the Cocktail Party*, *Springer Handbook of Auditory Research, 60*, edited by J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay (Springer, New York), pp. 75–109.

Kidd, G., Jr., Mason, C. R., Best, V., Roverud, E., Swaminathan, J., Jennings, T. R., Clayton, K., and Colburn, H. S. (**2019**). "Determining the energetic and informational components of speech-on-speech masking in listeners with sensorineural hearing loss," J. Acoust. Soc. Am. **145**(1), 440–457.

Kidd, G., Jr., Mason, C. R., and Gallun, F. J. (**2005**). "Combining energetic and informational masking for speech identification," J. Acoust. Soc. Am. **118**(2), 982–992.

Kidd, G., Jr., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., and Best, V. (**2016**). "Determining the energetic and informational components of speech-on-speech masking," J. Acoust. Soc. Am. **140**(1), 132–144.

Kryter, K. D. (**1962a**). "Methods for the calculation and use of the Articulation Index," J. Acoust. Soc. Am. **34**(11), 1689–1697.

Kryter, K. D. (**1962b**). "Validation of the Articulation Index," J. Acoust. Soc. Am. **34**(11), 1698–1702.

Li, N., and Loizou, P. C. (**2007**). "Factors influencing glimpsing of speech in noise," J. Acoust. Soc. Am. **122**(2), 1165–1172.

Li, N., and Loizou, P. C. (**2008**). "Effect of spectral resolution on the intelligibility of ideal binary masked speech," J. Acoust. Soc. Am. **123**(4), EL59–EL64.

Lutfi, R. A. (**1990**). "How much masking is informational masking?," J. Acoust. Soc. Am. **88**(6), 2607–2610.

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Lang. Cognit. Process. **27**(7-8), 953–978.

Miller, G. A., and Licklider, J. C. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**(2), 167–173.

Montazeri, V., and Assmann, P. F. (**2018**). "Constraints on ideal binary masking for the perception of spectrally-reduced speech," J. Acoust. Soc. Am. **144**(1), EL59–EL65.
Moore, B. C., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**(3), 750–753.
Olejnik, S., and Algina, J. (**2003**). "Generalized eta and omega squared statistics: Measures of effect size for some common research designs," Psychol. Methods **8**(4), 434–447.
Patterson, R. D., Robinson, K., McKeown, D., Zhang, C., and Allerhand, M. H. (**1992**). "Complex sounds and auditory images," in *Auditory Physiology and Perception*, edited by Y. Cazals, L. Demany, and K. Horner (Oxford Pergamon, New York), pp. 429–446.
Rennies, J., Best, V., Roverud, E., and Kidd, G., Jr. (**2019**). "Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort," Trends Hear. **23**, 1–21.
Roman, N., Wang, D., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**(4), 2236–2252.
Saberi, K., and Perrott, D. R. (**1999**). "Cognitive restoration of reversed speech," Nature **398**(6730), 760.
Schubotz, W., Brand, T., Kollmeier, B., and Ewert, S. D. (**2016**). "Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features," J. Acoust. Soc. Am. **140**(1), 524–540.
Shafiro, V., Sheft, S., and Risley, R. (**2011**). "Perception of interrupted speech: Effects of dual-rate gating on the intelligibility of words and sentences," J. Acoust. Soc. Am. **130**(4), 2076–2087.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304.
Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**(1), 318–326.
Stone, M. A., Füllgrabe, C., and Moore, B. C. (**2012**). "Notionally steady background noise acts primarily as a modulation masker of speech," J. Acoust. Soc. Am. **132**(1), 317–326.
Stone, M. A., and Moore, B. C. (**2014**). "On the near non-existence of 'pure' energetic masking release for speech," J. Acoust. Soc. Am. **135**(4), 1967–1977.
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. L., and Brown, G. J. (**1999**). "Separation of speech from interfering sounds based on oscillatory correlation," IEEE Trans. Neural Networks **10**(3), 684–697.
Wang, X., and Humes, L. E. (**2010**). "Factors influencing recognition of interrupted speech," J. Acoust. Soc. Am. **128**(4), 2100–2111.
Warren, R. M. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**(3917), 392–393.
Warren, R. M., and Obusek, C. J. (**1971**). "Speech perception and phonemic restorations," Percept. Psychophys. **9**(3), 358–362.
Watson, C. S. (**2005**). "Some comments on informational masking," Acta Acust. Acust. **91**(3), 502–512.
Zwicker, E. (**1961**). "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," J. Acoust. Soc. Am. **33**(2), 248.
Zwicker, E., and Scharf, B. (**1965**). "A model of loudness summation," Psychol. Rev. **72**(1), 3–26.