Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation

Douglas S. Brungart[a)]
Air Force Research Laboratory, Human Effectiveness Directorate, 2610 Seventh Street, Wright-Patterson AFB, Ohio 45433

Peter S. Chang[b)]
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210

Brian D. Simpson
Air Force Research Laboratory, Human Effectiveness Directorate, 2610 Seventh Street, Wright-Patterson AFB, Ohio 45433

DeLiang Wang
Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

(Received 2 January 2005; revised 25 August 2006; accepted 15 September 2006)

When a target speech signal is obscured by an interfering speech wave form, comprehension of the target message depends both on the successful detection of the energy from the target speech wave form and on the successful extraction and recognition of the spectro-temporal energy pattern of the target out of a background of acoustically similar masker sounds. This study attempted to isolate the effects that energetic masking, defined as the loss of detectable target information due to the spectral overlap of the target and masking signals, has on multitalker speech perception. This was achieved through the use of ideal time-frequency binary masks that retained those spectro-temporal regions of the acoustic mixture that were dominated by the target speech but eliminated those regions that were dominated by the interfering speech. The results suggest that energetic masking plays a relatively small role in the overall masking that occurs when speech is masked by interfering speech but a much more significant role when speech is masked by interfering noise. [DOI: 10.1121/1.2363929]

PACS number(s): 43.71.Gv, 43.66.Pn, 43.66.Rq [GDK] Pages: 4007–4018

I. INTRODUCTION

When a target speech signal is masked by one or more interfering voices, two related but distinct processes are required for the listener to successfully understand the message spoken by the target talker. First, the listener must be able to *detect* the acoustic energy present in the target voice. In general, a listener would only be able to reliably detect those portions of the target voice that occur in time and frequency regions where the target contains at least as much energy as the masker. The acoustic elements of the target that occur in regions that overlap in time and frequency with a more powerful masking sound would be effectively eliminated from the stimulus and thus are unable to contribute to the perception of the target utterance. In this paper, the loss of acoustic information from the target that is caused by this type of masking will be referred to as *energetic masking.*

In situations where no interfering sounds are present in the stimulus, detection alone is generally sufficient to allow the listener to understand a target utterance. If enough acoustic elements of the target talker are detectable in the stimulus, the listener should be able to determine the spectro-temporal energy pattern of the speech and use it to recognize the target utterance. When an interfering sound is present in the stimulus, however, the task becomes more difficult. The listener now detects acoustic elements from both the target and the interferer, and has to find some way to distinguish between the two in order to extract the spectro-temporal energy pattern of the target talker from the combined stimulus.

Of course, this two-stage model of speech perception is in many ways an oversimplification. In particular, it ignores the fact that some spectro-temporal regions of the stimulus will contain enough masker energy to *distort* the target signal, but not enough energy to overwhelm the target signal and render it undetectable. In such cases, it might be perfectly reasonable to expect the distortions caused by the noise to impair the listener's ability to identify those regions as part of the target signal and use them to recognize the target utterance. This raises the specter of whether the loss of performance caused by this kind of distortion should be categorized as *energetic masking*, because it is caused by direct spectral overlap of the target and masker, or whether it should be viewed as nonenergetic masking related to an inability to tease apart the target and masking portions of the stimulus. Either view may be equally valid, but for the purposes of this paper, we will restrict the use of the term "en-

[a)]Author to whom correspondence should be addressed; electronic mail: douglas.brungart@wpafb.af.mil
[b)]Electronic mail: Chang.549@osu.edu

ergetic masking" to the loss of information caused by an overwhelming masker, and exclude confusions caused by signal distortion from that definition. It is worth noting, however, that this distinction is a relatively minor one when the target signal is a spectrally sparse stimulus like speech, because in such cases the acoustic mixture will tend to be dominated either by the target or the masker in almost all the spectro-temporal regions of the stimulus.

Within this conceptual framework, an important question in understanding how listeners process complex auditory stimuli is the extent to which the ability to identify a target speech signal in the presence of a masker is dependent on the elimination of acoustic information about the target signal due to spectral and temporal overlap with the masker (i.e., energetic masking), and the extent to which performance in these situations is limited by the inability to correctly segregate and recognize the spectro-temporal pattern of the detectable acoustic elements of a target signal amid a background of confusingly similar masking sounds. In order to explore this issue, a number of researchers have attempted to develop stimuli that reproduce the potential target-masker confusions that can occur in a speech-on-speech masking task in a stimulus with no spectral overlap between the target and masking speech signals (and thus no significant opportunity for energetic masking to occur). For example, Spieth, Curtis, and Webster (1954) high-pass filtered one talker at 1600 Hz and low-pass filtered the other talker at 1600 Hz, thus creating a stimulus containing two independently intelligible speech signals with no spectral overlap. These stimuli reproduced many of the masking effects associated with normal multitalker stimuli, but presumably generated little or no energetic masking because there was no spectro-temporal overlap in the target and masking signals. This technique was greatly expanded by Arbogast *et al.* (2002), who used cochlear implant simulation software to divide the speech signal into 15 logarithmically spaced envelope-modulated sine waves, and randomly assigned eight of these bands to the target speech and six other bands to the masking speech. Again, this resulted in a stimulus that presumably produced little or no energetic masking but retained the potential target-masker confusions that would normally be present in multitalker speech.

The above studies have attempted to eliminate the energetic masking component that would ordinarily occur in speech-on-speech masking. An alternative approach is to develop a stimulus that approximates the effects of energetic masking but eliminates the potential nonenergetic target-masker confusions that can occur in ordinary multitalker stimuli. In order to be effective, such a stimulus has to account for the amplitude fluctuations that normally occur in a masking speech stimulus, because these fluctuations produce dips in the masking speech that allow listeners to obtain clear "glimpses" of the target speech even when the overall signal-to-noice ratio (SNR) is very unfavorable (Assmann and Summerfield, 2004; Cooke, 2005; Culling and Darwin, 1994; Miller and Licklider, 1950). The simplest approach to this problem is to amplitude modulate a continuous speech-spectrum-shaped noise with the overall envelope of a natural speech masker (Bronkhorst and Plomp, 1992; Brungart *et al.*, 2001; Festen and Plomp, 1990; Hawley *et al.*, 2004). However, this simplistic approach cannot account for the fact that a speech masker can fluctuate differently in different frequency bands, thus allowing the listener the opportunity to hear glimpses of the target in different frequency regions at different times (Buss *et al.*, 2004). Of course, it is possible to amplitude modulate a noise with different envelopes in different frequency bands (Festen and Plomp, 1990). However, signals of this type are known to become recognizable as intelligible speech when they contain more than a few independently modulated frequency bands (Shannon *et al.*, 1995), and, once this happens, there is a real possibility that the masking "noise" could include speech-like spectro-temporal patterns that could be confused with the target speech and thus result in a significant amount of nonenergetic masking.

We propose a new signal processing technique called "ideal time-frequency segregation" (ITFS) that can approximate the energetic masking effects produced by a natural speech masker across both time and frequency in a stimulus with no audible masking signal that could potentially be confused with the target speech. This approach, which is based on the "ideal binary mask" notion that has been used as a performance measure in computational auditory scene analysis (CASA), uses *a priori* information about the time-frequency (T-F) composition of the target and masking signals to eliminate just those spectral and temporal regions of the target signal that would ordinarily be rendered acoustically undetectable by the presence of a more intense masking sound. By eliminating these T-F regions, this procedure is intended to retain the loss of information that would normally occur due to the effects of energetic masking. At the same time, this procedure performs what could be viewed as ideal time-frequency segregation by separating the T-F regions of the stimulus that potentially contain information about the target speech (and thus would be expected to contribute to listener performance in a speech perception task) from those regions that would only contain acoustic information about the masking voice (and thus could potentially reduce performance by distracting the listener's attention away from the T-F regions associated with the target speech). Thus, the net effect of applying this ITFS procedure is to create a stimulus that approximates the effect that energetic masking would have for an "ideal" listener who is able to successfully identify and utilize all of the detectable target information in the stimulus.

The remainder of this paper is organized as follows. Section II describes the technical details of how the ITFS technique was implemented. Section III describes an experiment that applied this technique to multitalker stimuli. Section IV describes an experiment that extended this approach to speech in the presence of a noise masker. Section V describes how the ITFS technique can be used as a conceptual tool in a larger framework for quantifying the roles of energetic and nonenergetic masking in auditory perception. Finally, Sec. VI summarizes the main conclusions from our experiments.

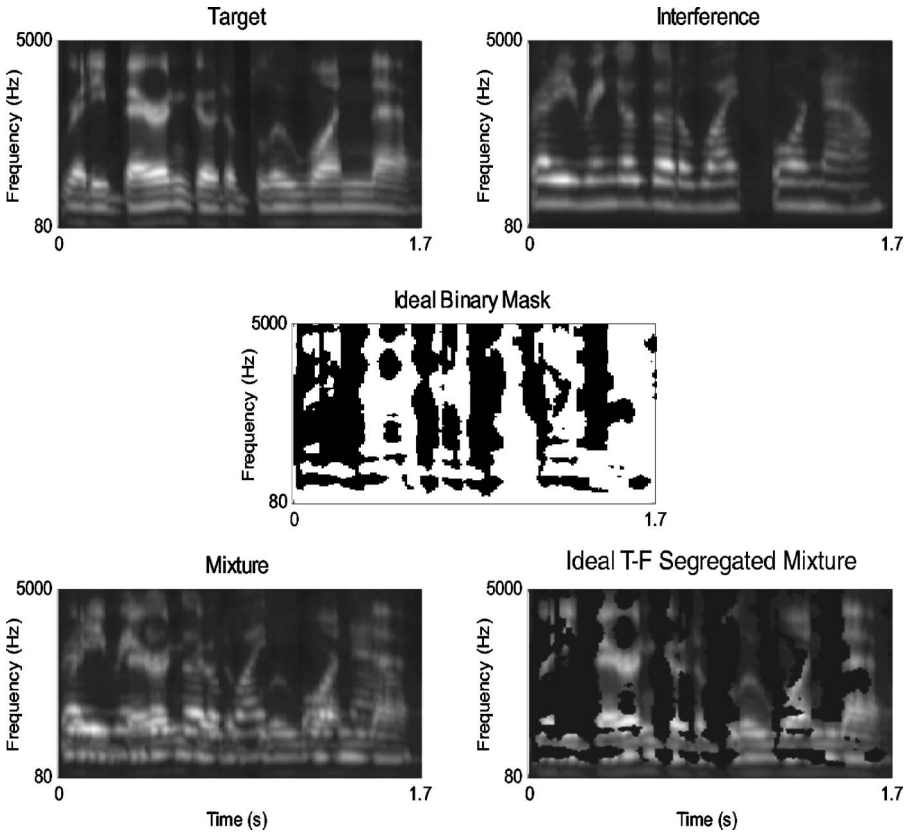

FIG. 1. An illustration of the ideal binary mask for a mixture of two utterances with equal overall rms levels. Top left: Two-dimensional T-F representation of a target male utterance ("Ready Baron go to blue one now"). The figure displays the rectified responses of the gamma-tone filter bank with 128 channels, where the energy value at each T-F unit is scaled to use the gray scale color map from black to white and is raised to the 1/4 power for better display. Top right: Corresponding representation of an interfering female utterance ("Ready Ringo go to white four now"). Middle: Ideal binary mask generated at 0 dB LC, where white pixels indicate 1 and black pixels indicate 0. Bottom left: Corresponding representation of the mixture. Bottom right: Masked mixture using the ideal mask.

II. IDEAL TIME-FREQUENCY SEGREGATION

A. Relationship to "ideal binary masking"

The concept of ITFS is very closely related to the concept of the ideal binary mask in CASA. Ideal binary masking is a signal processing method that simply retains those T-F regions of a mixture where the "target" source is stronger and eliminates the T-F regions where "interfering" sources are stronger. Such processing is called ideal because the mask definition is based on the target and interfering signals before mixing; also the ideal binary mask is the optimal binary mask in terms of SNR gain (Hu and Wang, 2004; Ellis, 2006). In this context, the term "mask" refers not to the addition of an interfering stimulus, as it does in psychoacoustics, but rather to a filter that completely eliminates certain portions of a signal (those assigned to a "zero" value in the mask) while allowing others (those assigned to a "one" value in the mask) to pass through unimpeded. Binary masking is typically based on a two-dimensional T-F representation where the time dimension consists of a sequence of time frames and the frequency dimension consists of a bank of auditory filters (e.g., gammatone filters). Thus the basic element in the ideal binary mask paradigm is a T-F unit corresponding to a specific filter at a particular time frame, and the binary mask itself is a two-dimensional matrix where each element corresponds to a single T-F unit. Those T-F units where the mixture is dominated by the target are assigned a one in the binary mask, and those where the mixture is dominated by an interfering sound are assigned a zero.

The above discussion makes it clear that the ideal binary mask is generated by checking whether the SNR in each T-F unit is greater than 0 dB. We can extend the definition of an ideal binary mask by introducing a predefined local SNR criterion (LC) so that those T-F units where the SNR is greater than a predefined LC value are assigned one in the ideal binary mask, and all the other T-F units are assigned zero. The commonly used ideal binary mask in CASA then corresponds to a LC value of 0 dB.

Figure 1 illustrates the ideal binary mask using a 0 dB LC for a mixture of the male utterance "Ready Baron go to blue one now" and the female utterance "Ready Ringo go to white four now," where the male utterance is regarded as the target. The overall SNR of the mixture (measured from the rms energy in each utterance) is 0 dB. In the figure, the top left panel shows the T-F representation of the target utterance, the top right panel the representation of the interfering utterance, and the bottom left panel the representation of the mixture. The middle panel shows the ideal mask, where white and black indicate 1 and 0, respectively. The bottom right panel shows the masked mixture using the ideal mask. Note that the masked mixture is much more similar to the clean target than the original mixture.

Figure 2 illustrates the effect that varying LC value has on the ideal binary mask for the two-talker speech mixture shown in Fig. 1. The left and right panels show the ideal mask and resulting resynthesized mixture with the LC value set at −12 dB (top row), 0 dB (middle row), and +12 dB (bottom row). As can be seen from the figure, increasing the LC value makes the ideal binary mask more conservative by requiring a higher local SNR in order to retain a particular T-F unit and hence reduces the total number of T-F units retained.

Although very few psychoacoustic experiments have

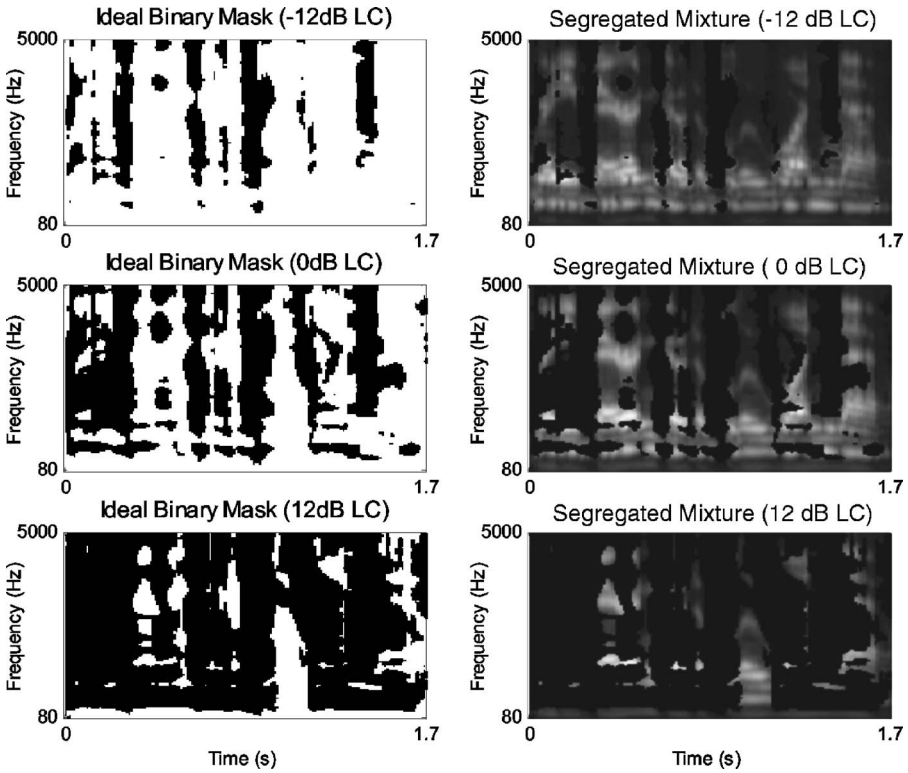

FIG. 2. An illustration of ideal binary masking at different LC values, using the same speech mixture of two utterances from Fig. 1. The three rows show representations for three different LC values (−12 dB, 0 dB, and +12 dB from top to bottom). The left column shows the ideal binary mask in each condition. The right column shows the corresponding masked mixtures using these ideal masks. Note that increasing the LC value makes the binary masking procedure more conservative and thus decreases the number of retained T-F units.

been conducted with ideal binary masks, the binary mask pardigm has been used extensively in CASA. The notion of the ideal binary mask was first proposed by Hu and Wang (2001) as a computational goal of CASA, and was further developed by Roman *et al.* (2003) and Hu and Wang (2004). Binary masks had been used as an output representation in the CASA literature (Brown and Cooke, 1994; Wang and Brown, 1999). Cooke *et al.* (2001) used the *a priori* mask—defined according to whether the mixture energy is within 3 dB of the target energy—in the context of robust speech recognition. Roman *et al.* (2003) conducted speech intelligibility tests and found that estimated masks that are very close to ideal ones yield substantial speech intelligibility improvements compared to unprocessed mixtures. Wang (2005) gave an extensive discussion on the use of the ideal binary mask as the computational goal of CASA.

B. Implementation

In applying the "binary mask" technique to auditory perception, we make the assumption that human listeners also perform a T-F analysis on the stimulus, and that the ideal listener would employ a segregation strategy similar to the one implemented in the ideal binary mask technique. In order to evaluate how real listeners might be able to perform a speech segregation task with such an "ideally segregated" signal, the technique described in the previous section was used to generate ITFS stimuli containing only those T-F regions of the stimulus that were locally dominated by the target speech. The procedures used to generate these stimuli were very similar to those used in previous studies that employed binary masks (e.g., Hu and Wang, 2004). For each stimulus presentation, the ITFS processing was based on a total of three input signals: a *target* signal, an *interfering* signal, and a *mixture* signal of the target and the interference. Each of these was first processed through a bank of 128 fourth-order gammatone filters with overlapping passbands (Patterson *et al.*, 1988) and with center frequencies ranging from 80 to 5000 Hz on an approximately logarithmic scale. The gammatone filter bank effectively decomposed the signals into arrays of 128 narrowband signals, which were further divided into 20 ms time frames with 10 ms overlap in order to produce a matrix of T-F units for each of the input signals. This choice of the filterbank and time windowing is commonly used in speech analysis but it is by no means optimal—other choices are certainly possible. For example, a bank of 64 or even 32 gammatone filters covering the same frequency range has been previously used. Once T-F decomposition is done, within each of the T-F units, a comparison was made between the energy of the target and that of the interference. The resulting local SNR for each T-F unit was then compared to a predefined LC value to determine whether to retain the unit.

Once this binary mask was defined, the output was resynthesized from the mixture using the same method that was described by Weintraub (1985) (see also Brown and Cooke, 1994; Wang and Brown, 1999), which accounts for across-filter phase shifts introduced by the gammatone filter bank. In resynthesis, the binary mask was used to weight the filter outputs in individual frames and the weighted outputs were summed across all frequency channels to yield the resynthesized ITFS wave form.

III. EXPERIMENT 1: EFFECTS OF IDEAL TIME-FREQUENCY SEGREGATION (ITFS) ON SPEECH INTELLIGIBILITY WITH THE CRM TASK

The basic premise of the ITFS technique is that it approximates the signal that would be available to a "perfectly

segregating" listener who could correctly extract all of the T-F units containing useful information about the target speech and completely ignore all of the other extraneous T-F units in the stimulus. Thus, one would expect ITFS processing to produce a large improvement in performance in listening situations where the masker is qualitatively similar to, and thus potentially easily confused with, the target signal. Experiment 1 was designed to examine the effect of ITFS processing on the intelligibility of a stimulus that has been shown to be highly susceptible to these types of target-masker confusions; namely, a target speech signal from the coordinate response measure (CRM) (Brungart *et al.*, 2001) masked by one, two, or three CRM phrases spoken by identical masking talkers and presented at the same level as the target speech. In order to examine the impact of the LC value on performance with ITFS signals, the stimuli were processed with LC values ranging from −60 dB to +30 dB.

A. Methods

1. Listeners

Nine paid listeners participated in the experiment. All had normal hearing and their ages ranged from 18 to 54. Most had participated in previous auditory experiments, and all were familiarized with the CRM task prior to conducting this experiment.

2. Speech stimuli

The speech materials used in the experiment were derived from the publicly available CRM speech corpus for multitalker communications research (Bolia *et al.*, 2000). This corpus, which is based on a speech intelligibility test first developed by Moore (1981), consists of phrases of the form "Ready (call sign) go to (color) (number) now" spoken with all possible combinations of eight call signs ("Arrow," "Baron," "Charlie," "Eagle," "Hopper," "Laker," "Ringo," "Tiger"); four colors ("blue," "green," "red," "white"); and eight numbers (1–8). Thus, a typical utterance in the corpus would be "Ready Baron go to blue five now." Eight talkers—four males and four females—were used to record each of the 256 possible phrases, so a total of 2048 phrases are available in the corpus.

For each trial in the experiment, a total of three audio signals were randomly generated and stored for offline ITFS processing prior to their presentation to the listeners. The first audio signal (the "target" signal) consisted of a CRM phrase randomly selected from all the phrases in the corpus containing the target call sign "Baron." The second audio signal (the "interfering" signal) consisted of one, two, or three different phrases randomly selected from the CRM corpus that were spoken by the same talker used in the target phrase but contained call signs, color coordinates, and number coordinates that were different from the target phrase and different from each other. Each of these interfering phrases was scaled to have the same overall rms power as the target phrase, and then all of the interfering phrases were summed together to generate the overall interfering signal used for the binary mask processing. The third audio signal (the "mixture") was simply the sum of the target and interfering signals for that particular stimulus presentation.

Note that, although all of the individual masking talkers in the mixture were scaled to have the same rms power, the overall SNR was less than 0 dB in the conditions with more than one interfering talker. In previous papers, we have clarified this distinction by referring to the ratio of the target speech to each individual interfering talker as the target-to-masker ratio (TMR), and by referring to the ratio of the target talker to the combined interfering talkers as the overall SNR (Brungart *et al.*, 2001). Under this terminology, the mixture with two equal-level interfering talkers would have a TMR value of 0 dB and an SNR value of approximately −3 dB.

3. Ideal time-frequency segregation

Prior to the start of data collection, each set of three audio signals ("target," "interferer," and "mixture") was used to generate a single ITFS stimulus at a single predetermined LC value. A total of 29 different LC values, ranging from −60 dB to +30 dB in 3 dB increments, were tested in the experiment. In addition, an "unsegregated" condition was included where the stimuli were simply processed using the ITFS technique with a LC value of negative infinity (thus including all T-F units in the resynthesized mixture). This control condition was essentially equivalent to simply presenting the mixture to the listener, but it also captured any distortions that might have occurred during the analysis and resynthesis portions of the ITFS processing.

4. Procedure

The listeners participated in the experiment while seated at a control computer in one of three quiet listening rooms. On each trial, the speech stimulus was generated by a sound card in the control computer (Soundblaster Audigy) and presented to the listener diotically over headphones (Sennheiser HD-520). Then an eight-column, four-row array of colored digits corresponding to the response set of the CRM was displayed on the CRT, and the listener was instructed to use the mouse to select the colored digit corresponding to the color and number used in the target phrase containing the call sign Baron.

The trials were divided into blocks of 50, each taking approximately 5 min to complete. Each subject participated in 90 blocks for a total of 4500 trials per subject. These included 150 trial combinations for each of the 30 LC values (including the unsegregated condition) evenly divided among three talker conditions (two-talker, three-talker, and four-talker, corresponding to one-interferer, two-interferer, and three-interferer, respectively). The trials were also balanced to divide the eight target speakers as evenly as possible across the trials collected in each condition for each subject.

B. Results and discussion

Figure 3 shows the percentage of trials where the listeners correctly identified both the color and the number in the

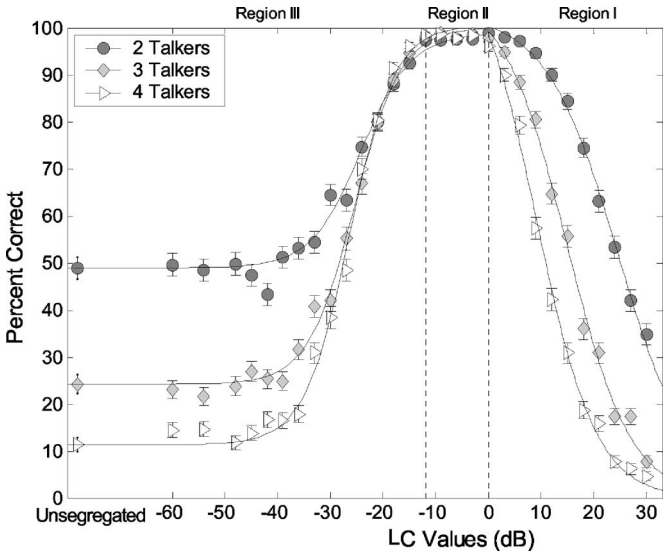

FIG. 3. Percentage of trials in Experiment 1 in which the listeners correctly identified both the color and number coordinates in the target phrase as a function of the LC values. The legend indicates the number of simultaneous talkers tested in the experiment. The error bars represent 95% confidence intervals (±1.96 standard errors) in each condition, calculated from the pooled data collected from all the subjects in the experiment. Because of the discontinuity in the performance curve at 0 dB, two different logistic curves were used to fit the positive and negative LC values in each talker condition (Cavallini, 1993). At negative LC values, this logistic curve was set to asymptote at the performance value achieved in the unsegregated control condition. Note that the target and maskers always were presented at the same overall rms level (i.e., 0 dB TMR) prior to the application of the ITFS procedure.

target phrase as a function of LC for each of two, three, and four simultaneous talker configurations in the experiment. The data were averaged across the listeners used in the experiment, and the error bars in the figure represent the 95% confidence interval of each data point. The points at the far left of the figure (labeled unsegregated) represent the control conditions where all of the T-F units were retained in the resynthesized signal. These points indicate that the listeners were able to correctly identify the color and number coordinates in the stimulus in approximately 50% of the trials with two simultaneous talkers, 25% of the trials with three simultaneous talkers, and 12% of the trials with four simultaneous talkers. These results are consistent with previous experiments that have examined performance in the CRM task with two, three, or four identical talkers (Brungart *et al.*, 2001).

Although the overall number of correct color and number identifications clearly decreased as the number of talkers in the stimulus increased, the general pattern of performance was similar for all three of the talker conditions tested. For purposes of discussion, it is easiest to divide these performance curves into three distinct regions that are clearly defined in all three of these performance curves shown in Fig. 3 and have very different interpretations with respect to the effect of LC. Before discussing each of these regions in detail, we describe the performance for the LC value of 0 dB, which corresponds the standard definition of the ideal binary mask discussed in Sec. II.

1. LC Value of 0 dB: Effects of energetic masking at a TMR value of 0 dB

As discussed earlier, the basic premise of the ITFS technique is that its application with a LC value at or near 0 dB preserves the effects of energetic masking in the stimulus but eliminates errors due to target-masker confusions. Under this premise, the point with 0 dB LC in Fig. 3 can be roughly interpreted as the theoretical maximum level of performance that could be achieved if the listener were able to successfully segregate and identify all of the detectable acoustic elements of the target talker in the stimulus.

An examination of the results in Fig. 3 shows that the listeners in the 0 dB LC condition correctly identified the color and number in the target phrase nearly 100% of the time even in the most difficult four-talker condition. This represents a dramatic improvement in performance over the unsegregated conditions, where performance ranged from 12% correct responses in the four-talker condition to 50% correct responses in the two-talker condition. Since the elimination of the T-F regions with negative local SNR values would not provide additional information about the target, our explanation for the large improvement in performance in the ITFS condition is that performance in the unsegregated condition was primarily dominated by nonenergetic masking effects due to the listener confusing the target and masking voices.

2. Region I: Energetic masking effects at LC values greater than 0 dB

The curves in Region I of Fig. 3 show that performance in the ITFS condition systematically decreased as the LC value increased above 0 dB. The fact that performance was near 100% when LC=0 dB and that it began to decrease almost immediately when LC increased above 0 dB is informative, because it implies that the listeners were able to obtain a significant amount of information from T-F regions of the stimulus that had local SNR values between 0 and +3 dB. This clearly shows that listeners can (and do) extract information from T-F regions of the stimulus with local SNR values as low as 0 dB, thus supporting the somewhat arbitrary choice of 0 dB as the nominal local SNR point differentiating between those T-F units which provide the listener with useful information about the target and those that only provide useful information about the masker.

The drop off in performance when LC$>$0 dB is also informative in another way. By the definition of ITFS, each 1 dB increase in the LC value produces a 1 dB increase in the minimum local SNR value required for a given T-F unit to be retained. This means that each 1 dB increase in LC above 0 dB eliminates exactly the same T-F units from the stimulus that would be eliminated if LC remained unchanged but the level of the masker increased by 1 dB. Thus, to a first approximation, the level of performance achieved for an ITFS stimulus with a LC value of $+L$ dB is approximately equivalent to the level of performance that could be achieved with just the energetic component of speech on speech masking in the corresponding stimulus with the overall SNR reduced by L dB (this assumption will be tested in the next experiment). The Region I performance with positive LC

values could therefore be interpreted as an indicator of the effect of purely energetic masking on multitalker speech perception with the corresponding reduction of overall SNR. These results would then suggest that energetic masking has a remarkably small impact on the performance with the multitalker CRM task. At a TMR of 0 dB, performance was near 100% even for a stimulus containing four simultaneous same-talker speech signals, which corresponds to an overall SNR of approximately −4.8 dB. Even when the effective SNR value was reduced by 30 dB (LC=30 dB), performance was near 35%correct in the two-talker condition and better than chance (3%) even in the four-talker condition.

3. Region II: Plateau in performance at LC values from −12 to 0 dB

In the range of LC values between −12 and 0 dB, the listeners consistently exhibited near-perfect identification performance (100% correct responses) in all three of the talker configurations tested. In this region, the stimulus contained all of the available speech information about the target, but only those T-F regions of the masker that were slightly more intense than the target. These units with stronger masking energy did not significantly interfere with the recognition of the target speech, presumably either because there were not enough of them to be perceived as a competing speech signal or because they only occurred in T-F regions that already contained a significant amount of energy in the target signal.

4. Region III: Target-masker confusion at LC values less than −12 dB

When the LC values used to generate the ITFS stimuli fell below −12 dB, the stimulus started to include progressively more T-F units in places where relatively little or no energy was present in the target. These units produced a rapid decrease in performance, from near 100% at LC value of −12 dB to approximately the same level of performance achieved in the unsegregated condition when LC=−40 dB. Decreasing the LC value below −40 dB caused no further degradation in performance, presumably because all of the relevant phonetic information in the interferer was already present in the stimulus when the LC value was −40 dB.

Decreasing the LC value below 0 dB has essentially no impact on the total amount of phonetic information in the target that is available to the listener. All of the T-F units that include usable information about the target are presumably included in the stimulus at a LC value of 0 dB. What happens when LC is reduced below 0 dB is that some of the T-F regions that primarily include phonetic information about the *interferer* are added back into the stimulus. Thus, the decrease in performance that occurs at negative LC could only be attributed to the listener becoming confused about which acoustic elements belong to the target and which acoustic elements belong to the interferer.

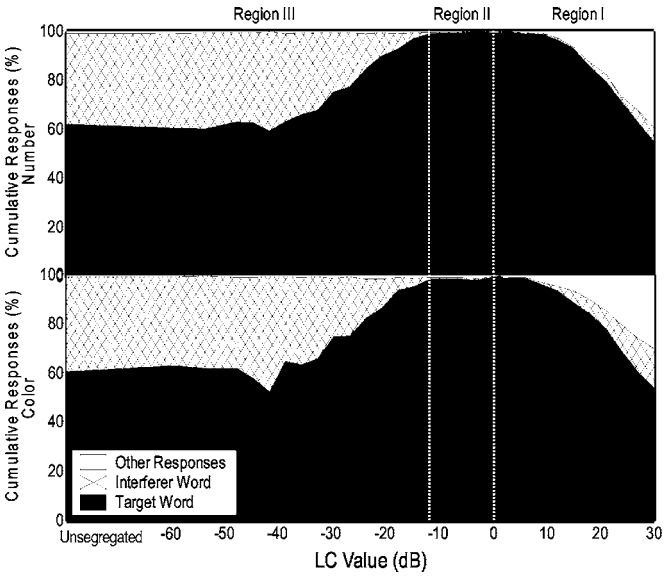

FIG. 4. Distribution of listener color and number responses in the two-talker condition of Experiment 1. The top panel shows the distribution of listener number responses in the experiment: the black area indicates correct responses that matched the number word in the target phrase; the mesh area indicates incorrect responses that matched the number word in the masking phrase; the white area indicates responses that did not match either of the number words contained in the stimulus. The bottom panel shows the same information for the color responses in Experiment 1.

5. Error analysis

Although the error rates shown in Fig. 3 provide an overall picture of the effect that the LC value of the binary mask had on performance in the CRM task, additional insights can be obtained by further analyzing the *types* of errors the listeners made in the experiment. The area graphs in Fig. 4 divide the listeners' color and number responses in the two-talker condition of the experiment into three categories: correct responses that matched the color or number word in the target phrase (black region in the graph), incorrect responses that matched the color or number word in the interfering phrases (meshed region in the graph), and incorrect responses that did not match either of the color or number words present in the stimulus (white region in the graph). Again, the results are somewhat different in the three regions of the figure. In Region II, where the LC value was between −12 and 0 dB, the listeners responded correctly nearly 100% of the time and no meaningful error analysis is possible. In Region I, where the LC value was greater than 0 dB, the incorrect responses were essentially random: the incorrect number responses matched the number in the masking phrase roughly 1/7 of the time, and the incorrect color responses matched the color in the masking phrase roughly 1/3 of the time. Again, this is consistent with the effects of energetic masking: the elimination of phonetic information from the target due to energetic masking has no particular tendency to bias the listener towards any alternative response word, so the incorrect responses are randomly distributed across all the possible alternatives available in the response set.

In Region III, where the LC value was less than −12 dB, the distribution of incorrect responses was much different. In

that region, virtually 100% of the incorrect responses matched the color or number contained in the masking phrase, and almost none contained a color or number word that was not present in the stimulus. This is consistent with the assumption that target-masker confusions are responsible for the decrease in performance in that condition: the addition of the acoustic elements dominated by the masking phrase at negative LC values introduced a second intelligible voice into the stimulus that confused the listener about which voice to respond to.

It is worth noting that there is never any point at any negative LC value where the listeners exhibited a significant number of incorrect responses that did not match the interfering phrase in the stimulus. At modestly negative LC values, one might expect that some point could occur where there would be enough low-level phonetic interferer elements in the stimulus to confuse the listener about the contents of the target phrase, but not enough to allow the listener to understand the contents of the interfering phrase. Such a point would be expected to produce an increase in overall error rate with a random distribution of errors similar to that seen in Region I of the figure. However, no such region exists at negative LC values in Fig. 3. This implies that the acoustic elements of the interferer that are added to the stimulus at negative LC values have no effect on the recognition of the target phrase until they themselves become intelligible and present the listener with an alternative interpretation of key words spoken by the target talker.

IV. EXPERIMENT 2: EFFECTS OF IDEAL TIME-FREQUENCY SEGREGATION ON SPEECH PERCEPTION IN NOISE

The results of Experiment 1 clearly show that application of the ITFS technique produces a dramatic improvement in performance in a multitalker listening task where the confusability of the target and masking signals plays a dominant role in determining overall performance. However, as a comparison it is also helpful to examine what effect the technique might have on the intelligibility of speech in noise, where confusability of the target and masking signals has a much smaller impact on performance than the spectral overlap of the target and masking signals. Also, there was a desire to test the validity of the assumption, outlined in Sec. III B 2, that there was a rough equivalence in terms of energetic masking between the information lost by a 1 dB increase in the LC value of the stimulus at a fixed SNR and a 1 dB decrease in the SNR of the stimulus at a fixed LC value. Thus a second experiment was conducted to examine the effect of ITFS processing on the perception of speech in noise.

A. Methods

1. Listeners

A total of nine paid listeners participated in Experiment 2. All had normal hearing and their ages ranged from 21 to 55. Most had participated in previous auditory experiments, and all were familiarized with the CRM task prior to conducting this experiment.

2. Stimulus generation

The target phrases used in the experiment were derived from the same CRM corpus used in Experiment 1. However, in Experiment 2 these phrases were masked by noise rather than speech. Two different types of Gaussian noise interferers were used to generate the stimuli used in the experiment. The first was a continuous speech-shaped noise masker that was spectrally shaped to match the average long-term spectrum of all of the phrases in the CRM Corpus (Brungart, 2001). The second was a speech-shaped masker that was modulated to match the overall envelope of a speech phrase that was randomly selected from all the nontarget phrases in the CRM corpus. This envelope was extracted by convolving the absolute value of the CRM phrase with a 7.2 ms rectangular window. The resulting envelope was then multiplied with a continuous speech-spectrum-shaped Gaussian noise to generate a modulated noise that simulated the amplitude fluctuations that typically occur in the overall envelope of a natural speech utterance (Brungart, 2001).

3. Ideal time-frequency segregation

Prior to the start of data collection, each of the 256 phrases that contained the call sign Baron was used to construct a total of 60 different stimulus wave forms for eventual presentation in the experiment. These stimulus wave forms consisted of all combinations of two different types of noise (continuous and modulated), ten different effective SNR conditions (ranging from −27 to 0 dB in 3 dB steps), and three different types of ITFS processing. These three types of processing were:

(a) Method 1 ITFS: Fixed mixture SNR, variable LC values. Method 1 used the same procedure for generating the ideal mask that was used in Experiment 1, where the LC value used to calculate the binary mask was varied and the resulting mask was applied to a stimulus with a fixed SNR value of 0 dB. Each Method 1 stimulus was generated by scaling the interfering noise wave form to have the same overall rms power as the target speech (i.e., an SNR of 0 dB), calculating the ideal mask for this mixture with LC set to one of ten values ranging from 0 to 27 dB, and using this ideal mask to resynthesize the speech from the 0 dB SNR mixture.

(b) Method 2 ITFS: Variable mixture SNR values, fixed LC value. Method 2 was designed to test the assumption, outlined in Section III B 2, that each 1 dB increase in LC value was roughly equivalent in terms of the effects of energetic masking to a 1 dB decrease in the overall SNR of the stimulus. Each Method 2 stimulus was generated by scaling the rms power of the interfering noise wave form to one of 10 SNR values relative to the target speech (ranging from 0 to −27 dB in 3 dB steps), calculating the ideal mask for this mixture with LC set to 0 dB, and using this ideal mask to resynthesize the speech from the same SNR mixture used to generate the binary mask. Note that this processing resulted in a stimulus with the same binary mask as the corresponding stimulus generated by Method 1 (i.e., the same set of T-F units retained) but a lower local SNR value within each retained T-F unit.

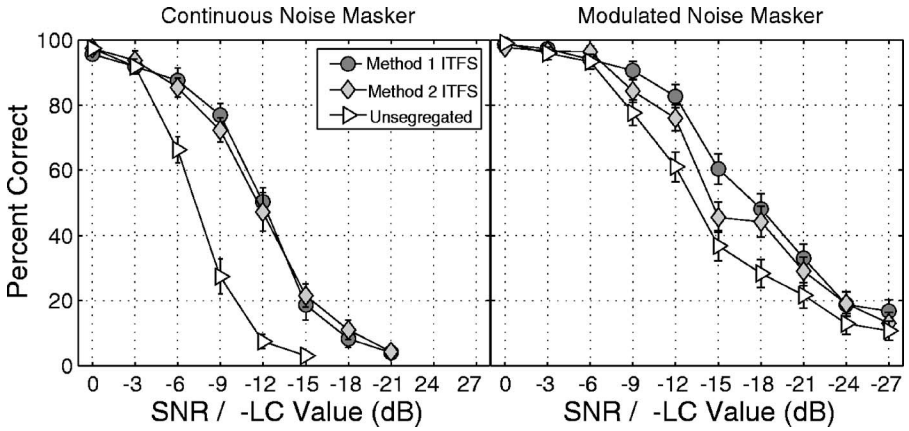

FIG. 5. Percentage of correct identifications of both color and number in each condition of Experiment 2. The left panel shows performance in the continuous noise conditions, and the right panel shows performance in the modulated noise conditions. The circles show performance in the Method 1 ITFS condition as a function of the negative of the LC value used to generate the stimulus. The diamonds in the figure show performance in the Method 2 ITFS condition as a function of the SNR value of the auditory mixture used to generate the stimulus. The open triangles show performance in the unsegregated control condition as a function of the SNR of the target speech. The error bars represent 95% confidence intervals (±1.96 standard errors) in each condition, calculated from the pooled data collected from all the subjects in the experiment.

(c) Unsegregated. Each unsegregated stimulus was generated by scaling the rms power of the interfering speech wave form to the appropriate SNR value relative to the target speech and resynthesizing the resulting mixed speech with a binary mask set to a value of 1 in every T-F unit. This processing was used to ensure that any artifacts introduced by the ideal mask processing software were also included in the unsegregated control condition.

4. Procedure

We used the same procedure as in Experiment 1 except for the following. In stimuli with SNR less than 18 dB, these stimuli were scaled to present the target at the same overall sound level [roughly 60 dB sound pressure level (SPL)]. In those stimuli with a SNR value greater than 18 dB, the stimuli were scaled to ensure that the overall level of the unsegregated stimulus (noise plus speech) would not exceed 80 dB SPL. The trials were divided into blocks of 50–100 trials, with each trial taking approximately 5 min to complete. Preliminary testing revealed that target speech was completely inaudible in the unsegregated continuous noise conditions with SNR lower than −15 dB and in the ITFS continuous noise conditions with SNR values lower than −21 dB (or equivalently those with LC values greater than 15 or 21 dB in Method 1). These eight conditions were eliminated from subsequent data collection. Thus, each listener participated in a total of 48 trials in each of 52 remaining stimulus conditions, for a grand total of 2496 trials for each of the nine listeners in the experiment.

B. Results and discussion

Figure 5 shows the percentage of correct color and number identifications in each condition of Experiment 2. The left panel shows performance in the continuous noise conditions, and the right panel shows performance in the modulated noise conditions. The circles show performance in the Method 1 ITFS condition as a function of the negative of the LC value used to generate the stimulus. The diamonds show performance in the Method 2 ITFS condition as a function of the overall SNR value of the mixture used to generate the stimulus. The open triangles show performance in the unsegregated control condition as a function of SNR.

Overall, the results in the unsegregated control condition were consistent with earlier experiments that have examined CRM performance with similar types of masking noise (Brungart, 2001). In both the continuous and modulated noise conditions, performance was near 100% when the SNR was near 0 dB. At lower SNR values, both conditions exhibited a decrease in performance with the ogive-shaped curve that typically occurs in speech-in-noise intelligibility tests. However, this decrease in performance was much more rapid in the continuous noise condition than in the modulated noise condition. In fact, performance in the unsegregated modulated noise condition was significantly better than chance (3%) even at the lowest SNR value tested. This result clearly illustrates the intelligibility advantages that can be obtained by listening to target speech "in the gaps" of a fluctuating masking sound.

Comparing the results from the unsegregated conditions of the experiment to those from the ITFS conditions, it is apparent that performance curves for the Method 1 and Method 2 stimuli were very similar to one another, and that they were almost identical in shape to those obtained in the unsegregated condition. This suggests that the application of an ideal binary mask with an LC value of 0 dB improved the intelligibility of a speech stimulus masked by noise by roughly 2–5 dB. While not inconsiderable, this improvement appears small relative to the dramatic 50 to 90 percentage point improvements obtained by applying the same binary mask to the 2, 3, and 4-Talker stimuli tested in Experiment 1. Thus it seems that the application of the ideal binary mask had less effect when the target speech signal was masked by noise than when it was masked by speech. These results also have two additional major implications.

The first implication is that there is a very strong (nearly one-to-one) relationship between the amount of masking caused by a 1 dB decrease in the SNR of a noise-masked

speech signal and the loss of acoustic information caused by a 1 dB increase in the LC value used to create an ITFS stimulus.

The second major implication of Experiment 2 is that the pattern of the T-F units preserved by the application of ITFS has a much greater impact on performance than the underlying local SNR values of these T-F units. In this experiment, the Method 1 and Method 2 stimuli were designed to produce exactly the same ideal masks, and thus to retain exactly the same T-F units, at each stimulus SNR. The only difference between the two methods was that the local SNR value within each retained T-F unit of the Method 2 stimulus was substantially lower than in the corresponding T-F unit of the Method 1 condition. Yet, despite these substantially higher local SNR values in the Method 1 condition, the results in Fig. 5 show that performance was nearly identical for the two methods at all SNR values tested. This result suggests that speech perception is much more limited by the listener's ability to determine where the energy in the target speech is located in the T-F domain than by the ability to extract specific target information within individual T-F units.

V. GENERAL DISCUSSION

A. Relationship between ITFS, speech segregation, and informational masking

Speech segregation is an extremely complicated process that depends on many variables. In the introduction, we suggested that a possible way to approximate speech segregation at a basic level is to view it as a relatively simple two-stage analysis of the T-F representation of the auditory mixture. The input to this analysis is provided by the auditory periphery, which can be viewed as a T-F analyzer with resolution that is limited in frequency by the bandwidths (critical bands) of the cochlear filters and in time by the occurrence of forward and backward masking within each critical band. Thus, the output from the auditory periphery could be viewed as an array of individual T-F units, with each unit representing the auditory signal that occurs at a particular time and frequency in the acoustic stimulus, and with the size (i.e., bandwidth and length) of each unit representing the smallest auditory event capable of being individually resolved by the auditory periphery.

Given this array of T-F units, the goal of speech segregation is to find a way to extract target speech from an acoustic mixture containing competing sounds. This can be viewed as consisting of two distinct, but interrelated, stages. The first stage is performed at the level of individual T-F units. Because each T-F unit is, by definition, too small to allow the resolution of individual sounds within the same unit, there are really only a few possible outcomes for a given unit within a complex acoustic mixture. When the unit contains substantially more energy from the target than from the masker, the characteristics of the T-F unit are essentially identical to those of the target alone, and there should be little loss of information due to the presence of the masker. When the unit contains substantially more energy from the masker, information from the target is expected to be lost, and the characteristics of the T-F unit should approximate those of the masker. When the unit contains comparable amounts of energy from the target and masker, the properties of the T-F unit would be corrupted, matching neither the target nor the masker. In the last two cases, where the masker energy in an individual T-F unit is comparable to or greater than the target energy in that unit, the resulting corruption or elimination of the target information would be an example of energetic masking.

In the second stage of the two-stage speech segregation process, the listener examines all of the T-F units in the mixture and uses the acoustic characteristics of each unit determined in the first stage, along with any available *a priori* information about the characteristics of the target, to determine which T-F units should be associated with the target and integrated into the single acoustic image of the target. This is a classic auditory grouping task, and it presumably makes use of a variety of grouping cues such as common onsets or offsets, amplitude or frequency co-modulations, common periodicity across frequency, and *a priori* templates for complex spectro-temporal patterns in speech or nonspeech sounds (Bregman, 1990).

Ideally, the output of the second stage would be a composite signal that includes all of the T-F units containing useful information about the target speech, and excludes all the other units. In reality, however, this segregation process is unlikely to be perfect, as suggested by the performance differences between an unsegregated mixture and the corresponding ITFS processed version shown in Fig. 3. Two kinds of errors could be made in this segregation process. Listeners could inadvertently exclude some T-F units that contain useful information about the target, resulting in a loss of relevant information from the target. They could also inadvertently include some T-F units that only contain acoustic energy from the masker, resulting in a garbled signal that might not be completely intelligible. We contend that these two types of segregation errors represent the very core of the nonenergetic form of masking, often referred to as *informational masking*, that occurs when the listener is unable to segregate the acoustically detectable portions of the target speech from the similar-sounding acoustically detectable portions of the interfering speech (Brungart, 2001; Cahart and Tillman, 1969; Freyman *et al.*, 1999; Kidd *et al.*, 1998; Pollack, 1975).

Because the underlying processes involved in speech segregation are extremely complicated, it is very difficult (or perhaps impossible) to completely isolate the contributions that energetic and informational masking make to overall segregation performance in realistic stimuli. However, if we are willing to make some simplifying assumptions, we could use a tool such as the ITFS technique to get a rough estimate of the effects that energetic and informational masking have on the perception of an arbitrary stimulus. The first approximation is that the application of the ITFS technique with a LC value of 0 dB provides a rough estimate of the effect that purely energetic masking has on the perception of a mixture. The second approximation is that the relative effect of informational masking on an arbitrary acoustic stimulus can be approximated by the smallest positive LC value required to

bring ITFS-processed recognition performance down to the level obtained in the unsegregated condition. This operational metric of informational masking is based on the first approximation plus the assumption that each 1 dB increase in the LC value eliminates the same T-F units (and thus produces approximately the same loss of information) as the energetic masking caused by each 1 dB decrease in the overall SNR value of the stimulus. By this metric, the amount of informational masking produced by the speech maskers in Experiment 1 ranges from 22 to 25 dB, while that produced by the noise maskers in Experiment 2 is from 3 to 5 dB. While this result is clearly consistent with the general notion that speech-on-speech masking produces far more informational masking than speech-in-noise masking, much more data will be needed to determine if this crude metric will prove to be a viable way to compare the effects of informational masking across different types of complex stimuli.

B. Caveats

We believe the ITFS technique can be a valuable tool for assessing the effects of energetic masking in complex speech perception tasks. Because it makes no assumptions about the characteristics of the underlying stimulus, the technique has the potential to be applied to other complex listening tasks that tend to produce informational masking. However, a number of caveats should be kept in mind when applying this technique.

The first is that the technique is much better suited for tasks that require the identification or recognition of acoustic stimuli than for those that require the detection of acoustic stimuli. What we have measured is the intelligibility (or recognition) of target speech in the presence of interfering speech or noise, not the detection of the target signal. In a detection task, there would of course never be any retained T-F unit in the target-absent stimuli, so the use of the ITFS technique would essentially be meaningless (that is to say that the detection level of the stimulus would be determined artificially by the LC value and not by the performance of the subject). Since we have not directly measured target detection, which is closely tied to energetic masking, one should keep in mind that our analysis on energetic masking is based on assumptions and approximations. The ITFS technique is intended to largely remove the effects of informational masking, and our data clearly show that such processing leads to dramatic improvement in target intelligibility despite energy overlap between target and interfering voices. Although we have argued that the technique should introduce minimal impact on energetic masking, future experiments are needed to test whether, or to what extent, our argument holds.

The second important caveat about ITFS technique is that it is intended only as tool to evaluate the relative importance of spectral overlap in the identification or recognition of complex stimuli. It makes assumptions about the segregation of speech stimuli that are clearly not true for human listeners, and it should not be taken as a plausible model of human speech segregation.

The third point of caution about the ITFS technique is that its utility very much depends on how accurately the bandwidths of the auditory filters and the lengths of the temporal windows used to decompose the stimulus into T-F units correspond to the spectral and temporal resolution of the human auditory system. The T-F units used in this experiment are believed to be reasonable estimates of the effective spectral and temporal resolution of human listeners for speech stimuli, but they may not be appropriate for other types of stimuli, particularly those involving transients. Even for speech stimuli, our ITFS processing uses a common but fixed way of decomposing a signal into T-F units and we have not evaluated the effects different T-F resolutions may have on intelligibility performance. In addition, our ITFS processing does not take any nonlinear effects of masking into account, so it will almost certainly provide incorrect results for stimuli involving very high-level masking sounds. Also not incorporated is nonsimultaneous energetic masking which can occur either due to the upward or downward spread of masking in the frequency domain or due to forward or backward temporal masking. In theory, it should be possible to develop a more sophisticated ITFS technique that would take these factors into account and produce a more accurate estimate of the effects of energetic masking in a wider range of listening environments.

Finally, it is important to note that the relatively insignificant role of energetic masking that seemed to occur in this set of experiments was probably related to the relatively small response set used in the CRM corpus. Because they are restricted to four phonetically distinct color alternatives and eight phonetically distinct number alternatives, the key words in the CRM phrases are relatively easy to understand even in extremely noisy environments (Brungart *et al.*, 2001). Presumably, the effects of energetic masking would be substantially greater for other speech perception tests based on speech materials with phonetically similar response alternatives, such as the Modified Rhyme Test (House *et al.*, 1965), or with larger response sets of phonetically balanced words or nonsense syllables. However, a recent study by Roman *et al.* (2003) used unrestricted and semantically predictable sentences from the Bamford-Kowal-Bench corpus (Bench and Bamford, 1979) and found that the binary masks that are very close to ideal binary masks with 0 dB LC are very effective in removing interfering sounds (competing speech or babble noise) and improving speech intelligibility in low SNR conditions. Further research is needed to empirically quantify how much greater the impact of energetic masking is in multitalker listening tasks based on these more difficult speech perception tests.

VI. CONCLUSIONS

This paper has introduced "Ideal Time-Frequency Segregation" as a tool for evaluating the relative contributions that informational and energetic masking make to the overall perception of complex speech stimuli. In general, the results of the experiment are consistent with those of earlier experiments that have examined informational masking in multitalker speech perception. Specifically, they show that informational masking dominates multitalker perception with the Coordinate Response Measure corpus, and that energetic

masking has a relatively much greater impact on speech-in-noise masking than on speech-on-speech masking. Indeed, the effects of energetic masking caused by the speech maskers tested in this experiment were so small relative to those produced by equivalently powerful noise maskers that they suggest that spectral overlap may play only a relatively small role in most speech recognition tasks involving more than one simultaneous talker. Further ITFS tests with different types of speech materials will be needed to determine if this conclusion generalizes beyond the limited CRM corpus tested in this experiment.

Our results also have strong implications for CASA research, where, as discussed in Sec. II, the notion of ideal binary masking was first introduced as a computational goal of CASA. First, the results from our experiments confirm that the ideal binary mask is a very effective technique to improve human speech intelligibility performance in the presence of competing voices. Second, while the commonly used 0 dB LC corresponds to a particularly simple and intuitive comparison, its validity has not been systematically verified in speech intelligibility tests. Our results show that this is indeed a good choice to achieve intelligibility scores near 100%. On the other hand, the 0 dB LC is near the borderline of a performance plateau that centers at −6 dB for the conditions of one, two, and three competing talkers. Since CASA systems must estimate the ideal binary mask, errors in estimation have to be considered. This suggests that the LC value of −6 dB is actually a better criterion to aim for than the 0 dB LC, at least for improving human speech perception in multitalker environments.

ACKNOWLEDGMENTS

This research was supported in part by an AFRL grant via Veridian and an AFOSR grant.

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

Assmann, P., and Summerfield, Q. (**2004**). "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. A. Ainsworth, and A. N. Popper (Springer-Verlag, New York).

Bench, J., and Bamford, J. (**1979**). *Speech Hearing Tests and the Spoken Language of Hearing-Impaired Children* (Academic, London).

Bolia, R. W. N., Ericson, M., and Simpson, B. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT Press, Cambridge, MA).

Bronkhorst, A., and Plomp, R. (**1992**). "Effects of multiple speechlike maskers on binaural speech recognitions in normal and impaired listening," J. Acoust. Soc. Am. **92**, 3132–3139.

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

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

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

Buss, E., Hall, J., and Grose, J. (**2004**). "Spectral integration of synchronous and asynchronous cues to consonant identification," J. Acoust. Soc. Am. **115**, 2278–2285.

Cahart, R., and Tillman, T. (**1969**). "Perceptual masking in multiple sound backgrounds," J. Acoust. Soc. Am. **45**, 694–703.

Cavallini, F. (**1993**). "Fitting a logistic curve to data," Coll. Math. J. **24**, 247–253.

Cooke, M. (**2005**). "Making sense of everyday speech: A glimpsing account," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 305–314.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Commun. **34**, 267–285.

Culling, J., and Darwin, C. J. (**1994**). "Perception and computational separation of simultaneous vowels: Cues arising from low frequency beating," J. Acoust. Soc. Am. **95**, 1559–1569.

Ellis, D. P. W. (**2006**). "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, edited by D. L. Wang and G. J. Brown (IEEE Press/Wiley, New York) (in press).

Festen, J., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Freyman, R., Helfer, K., McCall, D., and Clifton, R. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3587.

Hawley, M., Litovsky, R., and Culling, J. (**2004**). "The benefit of binaural hearing in a cocktail party: Effects of location and type of interferer," J. Acoust. Soc. Am. **115**, 833–843.

House, A., Williams, C., Hecker, M., and Kryter, K. (**1965**). "Articulation testing methods: Consonantal differentiation with a closed response set," J. Acoust. Soc. Am. **37**, 158–166.

Hu, G., and Wang, D. L. (**2001**). "Speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 79–82.

Hu, G., and Wang, D. L. (**2004**). "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Netw. **15**, 1135–1150.

Kidd, G., Mason, C., Rohtla, T., and Deliwala, P. (**1998**). "Release from informational masking due to the spatial separation of sources in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **104**, 422–431.

Miller, G., and Licklider, J. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **20**, Supp. 1.

Moore, T. (**1981**). "Voice Communication Jamming Research," in *AGARD Conference Proceedings 331: Aural Communication in Aviation*, pp. 2:1–2:6. Neuilly-SurSeine, France.

Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (**1988**). SVOS final report, part B: Implementing a gammatone filterbank. Rep. 2341, MRC Applied Psychology Unit.

Pollack, I. (**1975**). "Auditory informational masking," J. Acoust. Soc. Am. **82**, Supp. 1.

Roman, N., Wang, D. L., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Spieth, W., Curtis, J. F., and Webster, J. C. (**1954**). "Responding to one of two simultaneous messages," J. Acoust. Soc. Am. **26**, 391–396.

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 Netw. **10**, 684–697.

Weintraub, M. (**1985**). "A theory and computational model of auditory monaural sound separation," Ph.D. Dissertation, Stanford University.