The Role of Upward Spread of Masking in the Ability to Benefit from Asynchronous Glimpsing of Masked Speech

Erol J. Ozmeral

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology.

Chapel Hill
2011

Approved by:

Joseph W. Hall III

Emily Buss

Peter C. Gordon

Neil Mulligan

Abstract

Erol J. Ozmeral: The Role of Upward Spread of Masking in the Ability to Benefit from Asynchronous Glimpsing of Masked Speech

(Under the direction of Joseph W. Hall III and Emily Buss)

Previous studies have used asynchronous amplitude modulated (AM) maskers to investigate the ability of listeners to glimpse across frequency bands in a speech recognition task and achieve a release from masking when compared to unmodulated maskers (Howard-Jones and Rosen, 1993). In general, unmasking has been achieved only when frequency bands were spectrally wide. We hypothesize that previous work failed to show glimpsing for narrower bands due to the effects of upward spread of masking (USM) at the periphery. The current study contrasted previous masking conditions with a new method aimed at eliminating the deleterious effects of USM. Specifically, we presented the even and odd numbered bands of the asynchronous AM masker and target speech to opposite ears (dichotic stimulation). In experiment 1, results showed between 5 and 8 dB more masking release in the dichotic than the diotic asynchronous AM condition. Experiment 2 tested the effect of stimulus set-size on the ability to integrate across frequency bands. Results were comparable to experiment 1 for a closed-set task, but no benefit to dichotic asynchronous AM masking was observed in an open-set task. By listening to bands in the asynchronous AM masker dichotically, subjects were able to achieve greater unmasking with narrow frequency bands than previously had been shown. [Work supported by NIH R01 DC000418]

Acknowledgements

I would like to thank Drs. Joseph W. Hall III and Emily Buss for their direction and hours of counsel without which this project would have been impossible.

I would also like to thank the other committee members, Drs. Peter C. Gordon and Neil Mulligan for their guidance and constructive criticism during the review process.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AM: amplitude modulated

CMR: comodulation masking release

CNC: consonant-nucleus-consonant

ILD: interaural level difference

ITD: interaural time difference

MLD: masking level difference

SNR: signal-to-noise ratio

SPL: sound pressure level

SRM: spatial release from masking

SRT: speech reception threshold

USM: upward spread of masking

VCV: vowel-consonant-vowel

Chapter I

Introduction

Masking of speech signals

In everyday listening environments, following a conversation amidst numerous competing sounds can be complicated by target inaudibility and/or the confusions made by similar sounding sources. Traditionally, researchers have studied two forms of competition, or masking, of target speech signals. “Energetic masking” occurs when a masker overlaps with a target signal in both time and frequency, thereby causing the target to be inaudible. When a teacher must raise his or her voice above the sound of a classroom fan in order to be heard, this is an example of compensating for energetic masking. The second form of masking involves higher-level cognitive processing, which can occur if the masker is perceptually similar to the target or if the masker is presented contralateral to the target. Central masking, for example, can be the result of contralateral interference possibly related but not limited to confusions with the masking sound (Martin and Digiovanni, 1979). The classical example of informational masking is the confusions associated with understanding speech in the company of multiple other talkers (Cherry, 1953; Bronkhorst, 2000).

Synchronous amplitude modulated masking

In natural settings, such as a noisy city street or crowded bar, there is a combination of interfering sounds that fluctuate in time and frequency depending on their sources. Because most natural masking noises tend to vary in their spectro-temporal structure, listeners are

sometimes able to take advantage of the redundancy in speech across time and frequency by attending to regions in the signal which have the best signal to noise ratio (SNR; Miller and Licklider, 1950; Dirks and Bower, 1970; Howard-Jones and Rosen, 1993). Since steady-state noise is not the norm in natural settings, amplitude modulated (AM) maskers are often argued to reflect real-world scenarios. Often, these natural stimuli are comodulated (i.e., synchronous phase) across all frequencies (Nelken *et al.*, 1999). It is during the "off" phase of AM maskers that target speech has the best SNR. Taking advantage of the high SNR at the masker minima, also known as glimpsing (Li and Loizou, 2007; Gnansia *et al.*, 2008) or dip-listening (Peters *et al*., 1998), typically leads to improved identification, commonly measured by the speech reception threshold (SRT).

In one of the earliest studies on the effects of masker modulation on speech intelligibility, Miller and Licklider (1950) observed that speech intelligibility in the presence of a fluctuating masker was highly dependent on the rate of fluctuation. As the rate of modulation decreases below 200 Hz, intelligibility increases until around 10 Hz; however, as modulations are lowered below 10 Hz, entire words tend to be masked, and subsequently, performance declines. This finding has been shown to also depend on the type of speech material as well as the type of response measure employed (Buss *et al.*, 2009). In addition to studies that have found modulation rate to be an important parameter (Miller and Licklider, 1950; Buss *et al.*, 2009), the amount of masking release incurred by introducing masker AM can vary depending on the modulation depth (Gnansia *et al.*, 2008), as well as intensity level of the masker (Summers and Molis, 2004; George *et al.*, 2006).

Asynchronous amplitude modulated masking

Howard-Jones and Rosen (1993) tested the hypothesis that masking release associated with masker AM depends on the epochs of improved SNR coinciding across frequency. Their innovative design tested maskers which were separated into a given number of frequency channels, or bands, which were then amplitude modulated on and off at a rate of 10 Hz. Howard-Jones and Rosen controlled the phase of AM in neighboring bands, and that modulation was either in phase (synchronous) or 180 degrees out of phase (asynchronous). When the AM was out-of-phase in neighboring bands (i.e., asynchronously modulated), the masker resembled a checkerboard when viewed by its time-frequency representation, or spectrogram (Figure 1; Howard-Jones and Rosen, 1993). Figure 1 shows an example spectrogram of a 4-band asynchronous AM pink noise. The researchers tested vowel-consonant-vowel (VCV) identification in three primary masking conditions: unmodulated, synchronous AM, and asynchronous AM pink noise with varying numbers of frequency bands. Pink noise, which is similar to white noise but instead has equal power per octave, is sometimes used in studies of speech perception because its spectral shape roughly follows the long-term spectrum of speech. It was found that synchronous AM noise improved thresholds by 23 dB relative to the unmodulated noise condition; that is, there was a 23-dB masking release. In asynchronous AM conditions, there was some masking release when noise was filtered into 2 or 4 frequency bands -- 15.5 dB and 6 dB, respectively -- but close to zero unmasking was observed in the 8- or 16-band conditions. Interestingly, it was shown that thresholds in the 2-band asynchronous AM condition were significantly lower (i.e., better) than conditions in which one band was modulated and the other was left unmodulated. In other words, listeners were not performing well in the asynchronous modulation condition

based solely on information present in a subset of bands, but instead showed evidence of speech integration for signals that were unmasked asynchronously across time and frequency. This result was interpreted as showing evidence for "uncomodulated glimpsing." However, evidence of uncomodulated glimpsing did not occur in the 8 and 16 band cases.

It remains unclear why Howard-Jones and Rosen (1993) failed to find evidence of asynchronous glimpsing with greater than two bands. One possibility is that there is a perceptual limit on the ability to integrate asynchronous speech information when speech is distributed across a large number of frequency bands, but other evidence makes this unlikely (Buss *et al.*, 2004). In a speech identification experiment, Buss and colleagues (2004) determined masked identification thresholds for AM speech filtered into 2, 4, 8, or 16 frequency bands. Speech reception thresholds were determined for this modulated speech presented in a steady-state pink noise. Speech tokens were either synchronously or asynchronously modulated. Results of this study showed only a modest benefit for synchronous AM compared to asynchronous AM when the speech itself is modulated, and therefore, provided evidence for spectro-temporal integration of asynchronous speech information even when there are as many as 16 relatively narrow bands.

This finding -- that integration is possible for greater than 2 or 4 bands of asynchronously modulated speech -- prompted consideration of alternative explanations for Howard-Jones and Rosen's failure to find evidence in complimentary conditions where the noise was asynchronously modulated. One possible explanation for why synchronous AM noise had the largest masking release in Howard-Jones and Rosen's data is that better performance in the synchronous AM noise is aided by comodulated masking release (CMR; Hall *et al.*, 1984). In short, CMR is the improvement in detection thresholds seen when

comodulated off-frequency maskers are added to an on-frequency masked target. While CMR could have played some role in the results of Howard-Jones and Rosen, it is unlikely to account for synchronous/asynchronous AM differences on the order of 20 dB. Studies have shown CMR to have relatively small contributions to performance with suprathreshold stimuli, including speech (Grose and Hall, 1992; Hall *et al*., 1997; Kwon, 2002; Buss *et al*., 2003).

Another possibility is that better performance in the synchronous than asynchronous AM noise condition may be due to the negative effect of upward spread of masking associated with the asynchronous AM noise. Upward spread of masking (USM) is the phenomenon in which a masker positioned below the target in frequency will cause substantial energetic masking of that target. The amount of masking (in dB) is greatest when the masker is at a high intensity (Wegel and Lane, 1924). In the case of asynchronous AM masking, as described above, the advantage of selectively listening to unmasked frequency regions of target speech is likely to be reduced due to the USM from the lower frequency regions in which the masker is in the "on" phase of AM (Figure 2). That is, when an even-numbered frequency band is in the "off" phase of modulation, there is a neighboring odd-numbered band just below it which is "on" and contributing energetic masking. The same is true for each odd-numbered band, with the exception of the first, lowest frequency band. This effect is expected to be more detrimental when the frequency bands are narrow since any upward spread can mask a large proportion of the neighboring unmasked region. Hence, each masker has greater potential to degrade performance via USM when there are large numbers of bands, due to close proximity to neighboring speech bands.

Figure 2 shows the spectrograms of the output of a bank of auditory filters simulated by 128 gammatone filters (Slaney, 1998) for asynchronous AM noise separated into either 4 (left) or 8 (right) bands. With fewer, broader bands, USM does not affect the frequency regions associated with the noise in the "off" phase nearly as much as it may when there are more, narrow bands. It is clear from these plots that as the number of bands increases, excitation associated with the unmasked regions also increases. This is in accord with the results of as Howard-Jones and Rosen (1993), who showed less benefit of asynchronous AM masking for multiple, narrower masker bands.

Since listeners can integrate speech information distributed across a large number of asynchronous speech bands, as Buss and colleagues (2004) showed, Howard-Jones and Rosen may have shown only minimal integration because USM degraded the quality of the available speech. Importantly, Howard-Jones and Rosen presented diotic stimuli, meaning that all stimuli were presented to both ears symmetrically. Since USM occurs when asynchronous AM maskers are summed together at the periphery, it is expected that the effects of USM should be greatly diminished or eliminated if the even and odd numbered bands are presented to opposite ears. Figure 3 shows the spectrogram of just the even and just the odd bands for an 8-band asynchronous AM masker, as they would be represented by the auditory periphery in a dichotic presentation. So, by dividing the bands across the ears, the peaks of modulation will no longer exert USM on the dips of modulation in the higher neighboring band, leaving the listener a better opportunity to identify the speech.

Chapter II

Experiment 1

Experiment 1a

The first experiment adhered closely to the work by Howard-Jones and Rosen (1993), and included dichotic conditions, in which the even- and odd-numbered bands of stimuli were presented to opposite ears. This method was chosen because it should reduce the effect of USM at the periphery, which could underlie Howard-Jones and Rosen's failure to show uncomodulated glimpsing for asynchronous AM maskers with greater than two bands. The goal is to determine whether asynchronous glimpsing in the Howard-Jones and Rosen study was limited by USM, and whether the auditory system can indeed integrate asynchronous cues for speech identification across time and frequency with narrower spectral bands than seen before.

Method

Observers. Six native English speaking, young adults with no history of hearing loss or ear problems were recruited from the Chapel Hill community. All participants were screened for normal hearing, with a criterion of pure tone thresholds of 20 dB hearing level or better at octave frequencies from 250 to 8000 Hz in both ears (ANSI, 1996). No preference was made regarding sex or race.

Materials. The speech material was restricted to five tokens of each of 12 intervocalic consonants ([b d f g k m n p s t v z] as in [a*m*a]) spoken and recorded by an adult female speaker from this lab. Speech tokens were 523-664 ms, with a mean duration of 608 ms. Recordings were sampled at 44.1 kHz and digitally scaled to equal-rms level across tokens, then filtered into 2, 4, 8, or 16 frequency bands. Filter bandwidth was equivalent in logarithmic units, with bands spanning 0.1 to 10 kHz. Bands were generated using sixth order Butterworth band-pass filters. Speech tokens were up-sampled to 48828 Hz to conform to hardware specifications.

All masking noises were based on 0.1 to 10 kHz pink noise which, by definition, contains equal energy per octave band. Stimuli were generated digitally with duration equal to the longest possible speech token plus 300 ms (964 ms total duration), sampled at 48828 Hz. Modulated maskers were either modulated synchronously (Sync) or asynchronously (Async) across frequency, with a modulation rate of 10 Hz. To create these stimuli, first, pink noise was filtered using the same procedure discussed above for the speech stimuli. Second, each frequency band was modulated on and off at 10 Hz, with a starting phase alternating between starting on and starting off for consecutive bands (for example, see Figure 2). In order to limit spectral energy to the specified frequency region, 10-ms raised cosines were used to smooth these modulation transitions.

Maskers could be presented either monaurally (Ø), diotically (D), or dichotically (Δ). Monaural stimulation presents stimuli only to a single ear. Diotic stimulation presents the same bands to each ear, while dichotic stimulation presents the odd-numbered bands to the left ear and the even-numbered bands to the right ear. In all cases, speech bands were

presented with complimentary masker bands, and in some cases, only the masker bands were presented to an ear.

Design. An adaptive 'up-down' procedure was used to determine the speech reception threshold (SRT). The adaptive computer-controlled test procedure used a custom graphical user interface (GUI) administered through Matlab on a PC. Stimuli were presented through a pair of insert headphones (Etymotic ER-2) in a single-wall, sound-treated booth. The level of the speech was fixed at 45 dB sound pressure level (SPL) before filtering into bands, and no adjustment of the speech level was made to offset the overall energy reduction due to filtering. The initial masking level was set to 10 dB below pilot threshold levels determined for each condition. The level of the masking noise was turned up or down, depending on whether the previous response was correct or not. A correct response was followed by a trial in which the masker level increased by 4 dB, and an incorrect response was followed by a trial in which the masker was attenuated by 4 dB. The subject's estimated threshold was determined by computing the mean masker level at the last 24 of 26 track reversals. The test conditions were randomly arranged to avoid order effects. Each subject performed between three and four tests for each condition. The fourth estimate was measured if the first three thresholds were not all within 3 dB of each other. This occurred across all subjects for roughly 16 of the 21 conditions. Overall testing time was roughly 4 hours, typically spread out over three non-consecutive sessions.

Procedure. During the test, each speech token, randomly selected with replacement, was presented with the masker. Subjects responded by clicking a button on the GUI corresponding to the consonant heard out of a possible 12 consonants. In all, there were 21 test conditions. All thresholds were analyzed relative to the unmodulated noise condition

(Unmod). Two conditions used synchronous AM, one diotic and one dichotic (Sync-D and 8-Sync-Δ, respectively). For each asynchronous diotic and dichotic condition (Async-D and Async-Δ, respectively), stimuli were processed into 2, 4, 8, or 16 bands for a total of 8 additional test conditions. The key distinction between diotic and dichotic configurations is that the former has all stimulus bands presented to both ears, whereas the latter has just the even bands presented to the right ear and just the odd bands presented to the left ear.

Additionally, there were Async-Δ control conditions in which only the even or odd numbered speech bands were presented in the dichotic noise (Async-Δ-EVEN and Async-Δ-ODD, respectively) for a total of 8 dichotic controls. And finally, there were two controls for the Async-D condition filtered with 8 frequency bands and presented in a single ear (8-ODD-Ø and 8-EVEN-Ø). For reference, see Table 1.

Results and Discussion

Figure 4 shows the mean SRTs for each masker condition, expressed relative to the SRT for unmodulated pink noise. The SRTs for all conditions and bands are significantly different from the reference (Unmod) SRT (paired t-tests; $p < .05$) except for the 2- and 4-band Async-Δ-ODD conditions. Release from masking is greatest for the two Sync conditions (average of 23.8 dB), intermediate for the Async-Δ conditions (ranging from 22.2 to 14.4 dB as band number increases), and least for the Async-D conditions (ranging from 17.1 to 5.9 dB as band number increases). The difference in masking release between the two asynchronous conditions is between 5 and 8.5 dB, with greater masking release for the dichotic conditions. It is important to also note that the roughly 23-dB release from masking observed in the Sync conditions is the same as that found by Howard-Jones and Rosen (1993). While we *did* find release for the Async-D condition at all bands, finding less

masking as the number of bands increases was also similar to the trends reported by Howard-Jones and Rosen. A linear contrast in a one-way ANOVA with 4 levels of band confirmed this trend ($F[1,5]=201.9$, $p < .001$).

Control measures taken in the study are useful in assessing the possibility that a listener was simply attending to a subset of bands – either the even or the odd bands -- for the Async conditions, thereby not actually integrating across time *and* frequency. The 8-band, monaural controls (8-ODD-Ø and 8-EVEN-Ø) presented half of the stimuli to a single ear, whereas the dichotic controls (Async-Δ-ODD and Async-Δ-EVEN) only removed half of the speech bands and kept the alternate noise bands in the opposite ear intact. The data show that the monaural controls tend to have greater release from masking than their respective dichotic controls; for example, at 8 bands, release from masking in the 8-EVEN-Ø condition is 5.8 dB greater than the condition in which the contralateral noise is added (8-Async-Δ-EVEN). This may be related to the presence of noise in the opposite ear creating cross-ear interference, a possibility that will be addressed later on in the discussion.

In addition to the difference between the monaural and dichotic controls, it is important to point out the effect of adding additional speech information to the dichotic controls, as in the Async-Δ conditions. While adding only noise to the opposite ear creates a deficit in masking release compared to a monaural control, the inclusion of opposite-ear speech information increases release by 4.1 - 12.7 dB depending on the number of bands. In other words, there is more masking release in the Async-Δ conditions than in the dichotic controls for all numbers of bands. This difference was confirmed with a repeated measure ANOVA, including three levels of dichotic condition (Async-Δ, Async-Δ-EVEN, and Async-Δ-ODD) and four levels of band number (2, 4, 8, and 16). The analysis indicates

significant main effects of condition ($F[2, 10]= 49.97$; $p < .001$), band number ($F[3, 15]= 27.66$; $p < .001$), and a significant interaction ($F[6, 30] = 11.04$; $p < .001$). Combining the two dichotic controls to reflect only the better-case scenario (i.e., the maximum single-band release) still yields a significant main effect of condition ($F[1, 5]= 50.03$; $p =.001$) and band number ($F[3,15]= 39.90$; $p < .001$), but no interaction. Unlike Howard-Jones and Rosen's study, these results at greater than two bands suggest that brief epochs of improved SNR do not need to be synchronous across frequency in order to contribute to speech reception, and this will be examined further in the general discussion.

Information Transmission. One method of analyzing the difference between conditions is to calculate the amount of information transmitted to the listener on the basis of linguistic features of the stimuli (Miller and Nicely, 1955). Miller and Nicely identified five features that differentiate consonants in the English language. The features are based on articulation of the consonants and include: voicing, nasality, affrication, duration, and place of articulation. Based on information transfer analysis and treating each feature as a channel for information, Miller and Nicely break down a simple confusion matrix into smaller matrices based on each of the five separate channels. This kind of feature analysis can provide us with an understanding of which speech features transmit the most information towards accurate speech recognition, and also allow us to contrast speech information across conditions. For example, it is of interest to know whether information transmission differs between the presence of asynchronous dichotic and asynchronous diotic maskers.

A feature analysis was performed on the data of the current study to learn more about possible interactions between information content of the stimuli and the masker modulation manipulations. We were most interested in seeing whether there were effects beyond the

hypothesized elimination of USM and whether there were inherent transmission differences for speech masked by synchronous and asynchronous AM maskers. Percent transmission (in bits per stimulus) was calculated by dividing the number of transmitted bits by the total number of bits possible if no mistakes were made. Data were based on the individual error analyses of each listener (i.e., confusion matrices for each condition). This analysis was first performed on each condition to determine if information transmission was indeed dependent on the type of masker modulation presented to the listener. Table 2 shows the percent transmission results for each of the five features for each condition. High percentage indicates more efficient transmission of information for that feature. In general, percent transmission is highest for Nasal features and lowest for Affrication and Place features. Since the asynchronous diotic/dichotic manipulation is of particular interest, results from Table 2 are graphically shown in Figure 5 as a function of band number. As can be seen in Figure 5, the nasality feature transmits information more efficiently than any other feature for these conditions, meaning listeners do not make many errors in categorizing a stimulus as nasal (/m/ or /n/) compared to non-nasal (all other consonants).

Perhaps most importantly, the results of error analysis can show us the contribution to performance from each consonant. In other words, we can determine the trial-by-trial accuracy during a condition based on each individual trial and collapse across like consonants. Figure 6 shows the proportion of correct responses for each target consonant (i.e., the diagonal of a confusion matrix; e.g., see Table 3-5) plotted for each of the 8-bands Sync and Async conditions. A 3-factor repeated measures ANOVA with 2 levels of stimulation (diotic or dichotic), 2 levels of condition (Sync or Async), and 12 levels of consonant (all VCVs) showed no significant main effects of stimulation or condition, but a significant main

effect of consonant ($F[11,55]=14.9$, $p < .001$). This tells us that while each individual consonant contributed to performance unequally, this effect of consonant did not differ across stimulation and modulation conditions. If sphericity is not assumed, two-way interactions and the three-way interaction between factors failed to reach significance ($\alpha > 0.05$). From Figure 6, it is clear that not all consonants are the same in difficulty. For example, the nasal consonant /n/ was the consonant associated with the best identification. Furthermore, because there were no interactions between factors, varied difficulty among consonants did not affect individual conditions differently.

Upward spread of masking. The results of this study support the idea that the diotic asynchronous glimpsing observed by Howard-Jones and Rosen (1993) was limited by effects related to energetic masking at the auditory periphery. Masker bands in the "on"-phase likely introduced energetic masking into the neighboring spectral regions that were associated with the "off" –phase of modulation. This would be especially true for neighboring bands above the maskers in frequency (i.e., USM; Wegel and Lane, 1924). By presenting the alternating bands to opposite ears, the current study eliminated the effect of USM, and the result was between 5 and 8 dB of additional release from masking in the Async-Δ condition. Since this unmasking was greater than that associated with the control conditions, it is argued that listeners were integrating speech information across frequency and across ears, taking advantage of regions of high SNR distributed across frequency. This constitutes uncomodulated glimpsing.

Central masking. Another important point to consider is the effect that central masking may have had in the Async-Δ conditions and their controls. It is puzzling why masking release for the 8-Async-Δ and 16-Async-Δ conditions was smaller than for the

Sync-Δ conditions. Asynchronous masking elevated thresholds by 7.3 and 8.3 dB for the 8-Async-Δ and 16-Async-Δ conditions, respectively. The answer may lie in a form of central masking that may affect asynchronous AM masking independently of presentation type (e.g. dichotic or diotic). That is, contralateral maskers do not contribute to energetic masking, but they might introduce masking at a central level in the brain. This effect could be related to findings in the literature described as central masking (Martin *et al.*, 1965; Martin and Digiovanni, 1979) or informational masking (Zwislocki, 1971; Smith *et al.*, 2000).

In a study by Brungart and Simpson (2002) listeners were found to have greater difficulty identifying monaural speech when it was masked by a dichotic, speech competitor than when the competing speech was only in the ipsilateral ear. This effect disappeared when the contralateral ear (i.e. the opposite ear from the target speech) was presented with steady-state noise, indicating that the contralateral competition requires a signal qualitatively similar to the target to cause a disruption in speech segregation. While the present study did not use competing speech as maskers, the maskers were spectro-temporally more complex than steady-state or even synchronous AM noise. The data show that identifying speech with only half the bands presented to a single ear was less difficult in the monaural controls than in dichotic controls; therefore, it is possible that the addition of the contralateral, opposite-phase masker in the dichotic controls greatly reduced unmasking due to central effects. This can be seen in Figure 4, which shows the 8-band odd- and even-band monaural controls are associated with 2.9 and 5.8 dB more unmasking than the dichotic controls, respectively.

Furthermore, a masking effect attributable to across-ear interference may also have occurred due to miscuing. Such an effect may trigger the attention of the listener through a weighting process in which regions of masker minima may be given high weights due to

optimal SNR. Conversely, when maskers are at their highest level, i.e. low SNR, these weights may be reduced. In the current study, masker minima co-occur with masker maxima at other frequencies, which may lead to miscues for the listener's attention. Since this effect is not necessarily dependent on the presentation type (e.g. diotic or dichotic), it follows that performance in the diotic asynchronous AM conditions in the present study and in Howard-Jones and Rosen's study may also be detrimentally affect by miscuing.

Experiment 1b

Data in experiment 1a supported the conclusion that there was indeed an added advantage to dichotically presenting the asynchronous speech and maskers, such that masker bands presented to each ear were synchronously modulated, but that modulation was out of phase across ears. The data also supported the hypothesis that reducing the effect of USM allows listeners to take advantage of information present during epochs of advantageous SNR. However, an unforeseen percept was noted for the Async-Δ conditions; specifically, it was possible to perceive the masker as spatially separated from the target speech. Not all subjects reported perceived spatial separation, but when the percept was noticed, the target was heard as coming from a central location while the masker was perceived as two independent streams, one located at each ear. Without further experimentation, the role of perceived spatial separation in the release from masking in the dichotic conditions could not be fully evaluated.

Perceived spatial separation. While it has been generally shown that masked speech recognition can be aided by spatially separating the target speech sound from competing sources (Hirsh, 1950; Dirks and Wilson, 1969; Saberi *et al.*, 1991; Freyman *et al.*, 1999), the amount of spatial release from masking (SRM) tends to be less than 10 dB (Freyman *et al.*,

1999; Arbogast *et al.*, 2002). In general, greater improvements from spatially separating a target and noise are obtained when there is a breakdown in segregation or attention due to the similarity between the signal and masker (Brungart *et al.*, 2005; Edmonds and Culling, 2006). Conversely, the pink noise maskers used in experiment 1a are perceptually distinct from the target speech signals. Therefore, speech recognition in the present task is less likely to be affected by difficulties in sound source segregation. Furthermore, since the data show greater than 14 dB release from masking in the Async-Δ conditions, it is unlikely that such unmasking is solely due to the perceived spatial separation. In order to evaluate the possible role of SRM, however, a series of unmasking tests was conducted. The amount of unmasking associated with SRM was assessed in the absence of manipulations designed to reduce USM.

Masking-level difference. Locating the source of a sound is generally understood to involve two binaural cues: the interaural time difference (ITD), which is the difference in time of arrival of a sound at the two ears, and the interaural level difference (ILD), which is the signal level disparity between the ears (in dB). When a masker (N) and signal (S) have the same binaural characteristics, detection and identification can have thresholds noticeably higher than when N and S have different in binaural characteristics. This improvement in signal detection arising from binaural difference cues has been termed the masking-level difference (MLD; Hirsh, 1948; Moore, 2003). In the MLD paradigm, researchers determine the threshold for a masked signal in which both N and S have identical binaural characteristics, resulting in similar binaural percepts, and compare those thresholds to the case in which N and S have different binaural characteristics, resulting in different spatial percepts. MLDs can vary depending on the type of signal (e.g., speech) or masker (e.g., pink noise) and their respective spectral and temporal characteristics (e.g. modulation depth,

bandwidth, intensity; see Moore, 2003). The current experiment determined the MLD for masking conditions tested in experiment 1a in order to simulate possible contributions of SRM in the Async-Δ conditions.

Method

Observers. Six observers were recruited in experiment 1b, all meeting the inclusion criteria stated above. Three listeners had previously participated in experiment 1a, two were naïve listeners, and one was excluded from data analysis due to floor effects.

Design. For direct comparison to the first experiment, experiment 1b tested each of three modulation types (Unmod, Sync, and Async) from experiment 1a. The procedures and stimuli were identical to those described above with the following exceptions. For each modulation type, the reference stimuli were presented in a single ear (i.e., monaurally; abbreviated NmSm), and binaural cues were introduced with contralateral masking noise. This approach maintains the same conditions of USM in the signal ear, while also allowing the manipulation of the binaural cues. Two conditions with binaural cues available were assessed while keeping the speech signal monaural (Sm): 1) masker presented diotically (No), and 2) uncorrelated noise presented between the ears (Nu). The first binaural condition can have the percept of the masker centrally located while the signal is distinctly lateralized to one side, and the second condition can have a percept of masker non-discrete in position, yet recognizably separate from the clearly lateralized signal position.

Procedure. Subjects responded by clicking a button on the GUI corresponding to the consonant heard, selecting from among the 12 alternatives. In all, there were three modulation types (Unmod, Sync, and Async) with three spatial configurations (Nm, No, Nu),

for a total of 9 conditions. The Async conditions were filtered into 8 bands. Each subject performed at least three threshold estimates in each condition. In the event that a particular condition yielded greater than 3 dB variability for a given subject, that condition was run one additional time. This occurred on average for 6 of the 9 conditions. Overall testing time was roughly 2 hours spread out over two sessions.

Results and Discussion

MLD was calculated as the difference in threshold between the reference case without binaural cues (NmSm) and either of the two test cases with binaural cues (NoSm and NuSm). Figure 7 shows the mean MLDs for each of three modulation types. The MLD for the Sync NoSm condition is the only MLD to reach significance ($p=.014$) in a two-tailed student-t test, but when multiple comparisons Bonferroni correction is used, the value is no longer significant at the $\alpha = .05$ level. The Unmod and Async NoSm conditions also neared significant MLDs ($p=.086$ and $p=.061$, respectively). Interestingly, while the NoSm MLDs tend to be higher than the NuSm MLDs a two-way ANOVA with 3 levels of condition (Unmod, Sync, and Async) and 2 levels of binaural configuration (NoSm and NuSm) yielded no main effects of binaural configuration. There was also no main effect of condition and no interaction between the two.

Contribution of spatial release from masking. Masking level differences for recognition of speech stimuli in experiment 1b were found to be minimal for all modulation conditions. This is consistent with previous studies that have found *recognition* of speech stimuli to be less affected by binaural differences than speech *detection* in noise (Levitt and Rabiner, 1967b; a; Wilson *et al.*, 1982; Culling and Colburn, 2000). For example, Wilson and colleagues tested recognition of 36 spondaic words in two binaural configurations, NoSo and

NoSπ (where π indicates 180° out-of-phase between the ears). The mean MLD for these configurations was 7.2 dB, but ranged from 4.4 to 10.0 dB depending on the specific spondee. In the current study, the largest and only statistically significant MLD was found to be 3.4 dB in the Sync condition.

One explanation for MLD differences between the Wilson et al. (1982) study and the present one relates to the type of target speech used in the two studies. Specifically, spondees have been shown to have important cues at relatively low frequencies where the MLD tends to be large, whereas VCVs depend more upon high frequencies where the MLD has been found to be small (Carhart *et al.*, 1966). The current data along with previous work (e.g., Wilson *et al.*, 1982) suggest that perceived spatial separation between speech and noise simulated by interaural differences has a minimal contribution to intelligibility differences seen in the current study (compare a *maximum* MLD of 3.1 dB for the Async condition in experiment 1b to as much as 8.5-dB greater masking release in the Async-Δ condition compared to the Async-D condition in experiment 1a).

Summary of Experiment 1a and 1b

Howard-Jones and Rosen (1993) showed that uncomodulated glimpsing of speech in asynchoronous AM maskers is possible for small numbers of bands. The current study shows that presenting odd numbered bands to one ear and even numbered bands to the other ear improves the ability of the listener to identify the target speech. According to our hypothesis, this is a direct result of the elimination of USM from neighboring bands in the diotic presentation. Comparison of the Async-Δ condition and dichotic controls in experiment 1a shows the degree to which information is combined across ears. Experiment 1b shows that this improvement is well beyond the improvement expected by a best-case spatial unmasking

scenario. However, it should be noted that as the number of bands increases, and consequently the bandwidth of each band narrows, performance still decreases relative to the Sync conditions, which begs the question of what other constraints are placed on the listener when the masker is asynchronously modulated. By splitting the alternating bands to separate ears, factors such as central masking effects may counter the improvements seen by the elimination of USM.

Chapter III

Experiment 2

Rationale

Results from experiment 1a and 1b show evidence of asynchronous integration everywhere except in the 2-band case, where very good performance was obtained in the Async-Δ-EVEN condition. Additionally, release from masking relative to the unmodulated control condition was as much as 22.2 dB in the dichotic asynchronous condition, just slightly below the roughly 23-dB release for the two Sync conditions. Experiment 1b confirmed that any perceived spatial separation between the masker and signal in the dichotic conditions was insufficient to account for the large release from masking seen in experiment 1a. Also, since there was an additional benefit of having both sets of masked speech bands in the Async-Δ conditions over the dichotic controls -- 4.1 to 12.7 dB greater masking release, depending on the number of frequency bands – it was determined that listeners did not utilize just the bands presented to a single ear.

Of interest to the current study was the robustness of this effect when more speech information is required in order to make a correct response. The response set-size for speech identification can change the benefit of masker AM due to changes in the amount of detail needed to perform the task. In a study by Buss and colleagues (2009), masking release for words in synchronous AM noise was found to be different depending on the set-size of the

speech recognition task. When listeners were asked to identify a target word without constraints, masking release was smaller than when they were asked to select from among three alternatives: in one set of conditions, masking ranged from 8.7 dB (open-set) to 14.5 dB (closed-set) for coherent 10-Hz AM modulation. The authors argued that reducing constraints on the response alternatives increases the amount of information necessary to perform well on the task. It follows that if the set-size is manipulated for the identification tasks, listeners will have greater difficulty in the conditions with the least acoustic speech information.

Experiment 2 examined uncomodulated glimpsing as a function of set-size. It was expected that, as Buss et al.(2009) showed, Sync conditions would have less unmasking for an open-set identification task than a closed-set identification task. This may consequently limit the overall unmasking for Async conditions. However, due to the importance of speech redundancy in an open set task, it is possible that uncomodulated glimpsing will be associated with greater evidence of integration across time and frequency due to the insufficient information present in each subset of bands (just odd and just even). Therefore, it is unclear whether reduced masking release will obscure greater need for integrated information. However, the initial hypotheses are that the elimination of USM will produce a general benefit for dichotic stimulation compared to diotic stimulation, and more evidence of glimpsing may occur in the open-set task than the closed-set task.

Method

Observers. Ten observers participated in experiment 2, and all met inclusion criteria stated above. Five participants were tested in the open-set protocol and five in the closed-set protocol. Seven of ten participants had been tested on one or both of the previous two experiments.

Materials. The speech material for experiment 2 was a set of 500 CNC words (Peterson and Lehiste, 1962), spoken by an adult male with an American accent. Recordings were 444-992 ms, with a mean duration of 744 ms. The sampling rate was 24414 Hz, and all signals were passed through an 8-kHz second order Butterworth low-pass filter. Recordings were digitally scaled to equal-rms level across tokens. Speech tokens were up-sampled to 48828 Hz to conform to hardware specifications. Filtering the speech into frequency bands (2, 4, 8, and 16) was performed using the same methods described above for experiment 1a.

All masking stimuli were identical to those in the experiment 1a with the exception that stimuli were generated digitally, with duration equal to the longest possible speech token plus 300 ms (1,292 ms total duration), sampled at 48828 Hz. All stimuli could be presented monaurally (Ø), diotically (D) or dichotically (Δ). Dichotic stimulation presented the odd-numbered bands to the left ear and the even-numbered bands to the right ear.

Design. An adaptive '2-up-1-down' procedure was used to determine the SRT. The same hardware, data collection, target sound level, masker level step size, and listening environment were used as in the first two experiments. The subject's estimated SRT was computed as the mean masker level at the last of 10 of 12 response reversals, and test conditions were randomly arranged to avoid order effects.

For this experiment, two protocols were employed. The first protocol was a 4-alternative-forced choice identification (*closed-ID*) task. Listeners responded by clicking a button corresponding to the presented CNC word from a display of four choices. The second procedure, a free response identification (*open-ID*) task, allowed the listener to respond by repeating the target word aloud; at that point the listener was visually presented with the

correct response and prompted to score his or her response as correct or incorrect using buttons displayed on the computer screen. An experimenter monitored the experimental session, including spot checks for correct self-scoring. As in experiment 1a, there were 21 experimental conditions: 1 reference condition (Unmod), 2 synchronous AM conditions (Sync-D and Sync-Δ), and 2 test conditions (Async-D and Async-Δ) with either 2, 4, 8, or 16 bands. Dichotic controls were tested for each Async-Δ condition, and there were additional 8-band monaural controls (8-ODD-Ø and 8-EVEN-Ø). For reference, see Table 1. In the event that a particular condition yielded greater than 3 dB variability for a given subject, that condition was run one additional time. This occurred across all subjects for roughly 12 of the 21 conditions in the *closed-ID* task and 17 of the 21 conditions in the *open-ID* task. Overall testing time was roughly 4 hours per procedure, spread out over three separate 1-1.5 hour sessions.

Results and Discussion

Closed-set speech reception thresholds. Figure 8 shows the mean SRTs for each masker condition, expressed relative to the SRT for unmodulated pink noise. Release from masking is greatest for the Sync-D conditions (average of 22.8 dB), intermediate for Async-Δ conditions (ranging from 23.2 to 18.4 dB release as band number increases), and least for the Async-D conditions (ranging from 17.7 to 5.1 dB as band number increases). Once again, as Howard-Jones and Rosen (1993) observed, an increase in band number in the Async-D conditions reduced the overall performance relative to the synchronous conditions. Submitting the Async-D and Async-Δ thresholds to a two-way ANOVA with 2 levels of condition and 4 levels of bands confirms a main effect of condition ($F[1,4]=27.69$, $p < .01$), a main effect of bands ($F[3,12]=19.3$, $p < .001$), and a significant interaction ($F[3,12]=6.71$, p

< .01). The interaction can be seen by the increase in SRT for the Async-D conditions as band number increases, compared to the relatively flat function of thesholds for the Async-Δ conditions. Just as the data from experiment 1a suggest, there is a clear advantage in masking release for dichotic asynchronous AM maskers (Async-Δ) over the diotic counterparts (Async-D). The mean data show between 5.5 and 13.5 dB greater masking release in the Async-Δ conditions, and the main effect of condition was confirmed by the ANOVA.

Control measures were taken to rule out the possibility that a listener was attending to either the even or odd bands alone. The data show a clear benefit for dichotic asynchronous masker presentation over the dichotic controls. Specifically, SRTs for the Async-Δ condition averaged 6.9 dB less than Async-Δ-ODD and 8.4 dB less than Async-Δ-EVEN. By allowing the listener more speech information in the Async-Δ conditions, performance is better than when only the odd or even speech bands are present. This is evidence for integration across ears and frequency bands.

Open-set speech reception thresholds. Figure 9 shows the mean SRTs open-set data, expressed relative to the SRT for unmodulated pink noise. Symbols and line styles reflect the associated masker condition. Masking release for both Sync conditions was dramatically reduced for the open-set relative to the close-set protocols, with an average of 9.4 dB masking release. Consequently, it is not surprising that masking release in the Async conditions would also be significantly reduced when compared to the closed-set protocol. For the Async-D conditions, masking release ranged from 7.6 to 4.2 dB as the number of bands increased. In comparison, for the Async-Δ conditions, masking release ranged from 9.0 to 5.9 dB. A two-way ANOVA with 2 levels of condition and 4 levels of band confirms no main

effect of condition ($F[1,5]=1.8$, $p = .237$). Figure 10 shows the individual SRTs relative to the reference condition (Unmod) as a function of band number. While it was surprising that there was no significant difference between diotic and dichotic presentations, individual differences show that masking release varied greatly for most of the conditions. In particular, one subject (top left) consistently performed better in the Async-D conditions when compared to the dichotic counterpart.

Control conditions, which presented only half of the spectral speech bands alone (8-ODD-Ø and 8-EVEN-Ø) or with contralateral asynchronous noise (Async-Δ-ODD and Async-Δ-EVEN), had significantly higher SRTs than the reference condition (Unmod). This is consistent with the notion that a relatively difficult speech task such as open set word recognition requires a great deal of speech detail and redundancy (Buss *et al.*, 2009), which these controls certainly lack.

The magnitude of uncomodulated glimpsing was calculated as the difference between thresholds in the Async-Δ and the better of the two control conditions. Mean values of uncomodulated glimpsing are plotted as a function of the number of bands in Figure 11, with symbol and line-style reflecting response conditions. In open-set data, glimpsing ranges from 16.3 to 9.2 dB depending on band number (solid line). Contrast those numbers to the case for the closed-set protocol, in which glimpsing is calculated to range from only 7.8 to 5.8 dB (dotted line). A repeated measures ANOVA was performed to evaluate the effect of response protocol on the magnitude of uncomodulated glimpsing. There were 4 within-subjects levels of band and a between-subject factor of protocol, and results confirm a main effect of protocol ($F[1,8]=20.7$, $p < .005$). This indicates that there is greater evidence of integration across time and frequency in the open set than the closed set protocol. Therefore,

while unmasking as a whole is greatly reduced in the open-set protocol, the magnitude of uncomodulated glimpsing is significantly greater than in the closed-set protocol.

While this experiment did indicate a significant difference between diotic and dichotic asynchronous AM conditions in the closed-set task, no difference was observed in the open-set task. The explanation for this is likely to be based in the nature of USM (Wegel and Lane, 1924). It is documented that USM is level dependent, such that at higher intensities, target frequencies above a masker frequency are more detrimentally affected by the spread of excitation (Moore *et al.*, 1998). In this study, a target level was chosen to provide the highest level of masking without being uncomfortable for the listener. Since the Sync conditions were associated with the best performance (i.e., highest masker level), these conditions essentially dictated the target level, 45 dB SPL. Once this level was chosen in experiment 1a, it was kept constant for subsequent experiments for comparison purposes. As a result, in the open-set protocol of experiment 2 which had reduced unmasking, it is likely that masking levels did not reach high enough intensities to produces large effects of USM. Therefore, since the Async-D case was not likely to be substantially affected by USM, it is logical that separation between the ears (Async-Δ) would not have the beneficial outcome that it had in prior experiments in the study.

This interpretation was evaluated with pilot data for the 8-band conditions in the open-set task, in which the target level was increased by 10 dB. Mean results from 2 listeners, shown in Figure 12, confirmed an advantage in dichotic presentation of the asynchronous AM. That is, there was a nearly 10-dB difference between thresholds in the Async-D and Async-Δ conditions. This is consistent with the interpretation that the higher signal level is associated with greater USM in the Async-D, supporting greater benefit of

dichotic presentation. Therefore, if the target level were increased in a future experiment, we would expect greater effects of USM in the Async-D conditions and therefore more improvement with dichotic stimulation.

Lastly, integration was calculated as the difference between the SRTs in the Async-Δ conditions and the better of the two control conditions, Async-Δ-ODD or Async-Δ-EVEN. Uncomodulated glimpsing was found to be significantly greater than zeros for both the open and the closed set data, with mean values of 12 and 6 dB, respectively.

Chapter IV

General Discussion

To understand how listeners comprehend speech under noisy conditions, it is important to investigate the spectro-temporal structure of the interfering noise, or masker, as well as the amount of information necessary for adequate speech recognition. To simulate real-world listening situations in the lab, researchers commonly mask speech tokens with an AM broadband noise to determine the ability of a listener to integrate temporally discrete samples of speech information (Miller and Licklider, 1950; Wilson and Carhart, 1969; Festen and Plomp, 1990; Eisenberg *et al*., 1995; Bacon *et al*., 1998). The capacity to attend to and integrate information in epochs of high SNRs coincident with masker minima has been termed dip-listening or glimpsing (Gnansia *et al.*, 2008). This method can assess the advantages of having intermittent speech cues versus speech masked by continuous noise, but fails to capture natural listening environments in which speech is corrupted by noise fluctuating in both time *and* frequency. The present study used arguably more realistic maskers since both time and frequency were modulated, but one could still argue that asynchronous AM maskers are far from natural sounding.

Effect of dichotic stimulation

The current study built upon previous work by Howard-Jones and Rosen (1993) who tested the hypothesis that listeners can combine brief glimpses of speech in different frequency regions even if those glimpses do not occur synchronously in time. The authors

tested two types of AM maskers: noise was filtered into logarithmically distributed bands, and those bands were modulated either synchronously asynchronously. At 2 or 4 bands, consonant identification in the presence of asynchronous AM noise was relatively good compared to synchronous AM. However, performance in the asynchronous AM condition became much worse as band number increased. In addition, Buss and colleagues (2004) have shown modulating speech masked by a broadband noise resulted in similar thresholds whether the speech was modulated synchronously or asynchronously across frequency regions. Both of these studies together suggest that glimpsing is not exclusive to temporal dips that coincide across frequency, but also pertains to asynchronous glimpses distributed over frequency. What was puzzling from the Buss *et al*. study was why asynchronous modulated speech showed similar thresholds for all numbers of bands, whereas Howard-Jones and Rosen failed to show uncomodulated glimpsing beyond two bands.

One hypothesis regarding the apparent discrepancy between Buss *et al*. and Howard-Jones and Rosen is based on USM. As stated previously, a masking effect can occur at high intensities in which a lower frequency band energetically extends to higher frequencies. The overall result is masking of a target message which is not spectrally overlapping with the masker. In Howard-Jones and Rosen's study, USM may have disrupted masker minima, which are ideal regions for glimpsing. On the other hand, since the Buss et al. study employed a modulated speech stimulus in the presence of a steady noise, USM was not an issue because the noise was not modulated.

Because USM is a peripheral phenomenon, it was hypothesized that presenting even numbered bands to one ear and odd numbered bands to the other ear would improve performance. The first experiment of the present study tested 6 normal hearing listeners'

ability to identify vowel-consonant-vowel (VCV) speech tokens in the presence of synchronous or asynchronous AM maskers either presented diotically or dichotically. The study followed the Howard-Jones and Rosen (1993) study, with the additional manipulation of dichotic stimulus presentation. Control conditions were designed to include only half of the speech bands – either just the odd or just the even numbered bands – to look for evidence of integration of speech information across bands. In some of these control conditions the masker was also restricted to just even or just odd bands, and in other conditions the contralateral asynchronous masker was present. These conditions proved highly informative when compared to the asynchronous masking cases, and supported an estimate of the degree of uncomodulated glimpsing.

To summarize the results of the first experiment, listeners benefitted from the presence of the synchronous AM masker when comparing SRTs to the reference case (unmodulated broadband noise). As in the Howard-Jones and Rosen study, diotic presentation of the asynchronous AM masker had mixed results. For 2 bands, performance was good, but not as good as in the synchronous conditions. As band number increased, thresholds in the asynchronous diotic conditions rose. Our study separated the even and odd bands between the ears and showed significant improvement, but not a full recovery to the threshold seen in synchronous masking. Dichotic presentation of the asynchronous AM masker improved performance by 5 to 8 dB compared to the corresponding diotic conditions. As in the diotic conditions, asynchronous dichotic masking release declined as band number increased. Performance on control conditions confirmed the fact that listeners were indeed glimpsing information from both ears in the dichotic asynchronous AM conditions, so it is unclear what limited performance in these conditions from reaching levels seen for the

synchronous AM conditions. One possibility is that listeners had greater difficulty in the asynchronous condition because masker minima, or dips, in the even bands coincided with masker maxima, or peaks, in the odd bands and vice versa. Peaks in half of the masker bands may have discouraged the listener from attending to the dips in the other half of the bands by reducing the perceptual weights given to the information in the dips (Buus, 1985). This may have also played an important role in the Howard-Jones and Rosen (1993) study. Since our design does not allow an evaluation of the possible role of miscuing, it is unclear what role miscuing played in both diotic and dichotic asynchronous AM conditions.

Effect of spatial release from masking

In experiment 1a, presenting even and odd numbered bands to separate ears was found to be more beneficial than presenting all bands to each ear, but it was unclear whether this was entirely due to eliminating USM. The alternate hypothesis that was tested in experiment 1b was that perceived spatial separation between target and masker could account for the added benefit of dichotic presentation. In experiment 1b, an MLD paradigm was designed to simulate a best-case scenario of SRM for the three modulation conditions in experiment 1a (Unmod, Sync, and Async).

Masking noise was either presented monaurally (Nm) or it was presented binaurally, either perfectly correlated between the ears (No) or uncorrelated (Nu). Targets were always presented monaurally (Sm) in order to avoid changing effects related to USM. The monaural signal resulted in a segregation of signal and masker in the binaural masker conditions (NoSm and NuSm). The MLD was calculated as the difference between the NmSm thresholds and either the NoSm or NuSm thresholds. Data revealed that there was no more than 4 dB of MLD for any condition, meaning that perceived spatial separation between

target and masker did not yield large benefits. In contrast, data from experiment 1a had up to 8.5-dB greater masking release in the Async-Δ condition compared to the Async-D condition. Therefore, while it is possible that SRM contributed to the better performance in the Async-Δ than the Async-D conditions, the magnitude of SRM for these stimuli suggest that eliminating USM is more likely the cause of the better masking release in the dichotic condition in experiment 1a.

Effect of speech task set-size

Experiment 2 was designed to assess the nature of asynchronous AM masking as a function of speech identification task. Previous research has reported that the amount of unmasking is determined somewhat by the complexity of speech information required to perform a speech recognition task. Particularly, Buss and colleagues (2009) recently observed that manipulating the set-size of speech identification tasks can greatly alter the needed amount of the target signal to perform well. They argue that as set-size increases, meaning the number of possible choices rises, listeners require more speech information to identify the target. The researchers tested normal hearing listeners on speech identification of CNC words in the presence of either unmodulated noise or synchronous AM noise. Task set-size was manipulated between groups by either asking the subjects to identify the target word in a 3-alternative forced choice (3-AFC-ID) paradigm or to report back the word they heard as a free response, sometimes described as an open set identification (open-ID). At 10-Hz masker modulation, masking release in the 3-AFC-ID group was roughly 7 dB greater than in the open-ID group. This is consistent with our results in experiment 2, which show unmasking to be 13.4 dB greater in the closed-set task than the open-set task for synchronous AM maskers. However, the interpretation of this difference is confounded by the fact that

masker level was lower at threshold in the open-ID protocol due to the task's greater difficulty. So, while we did observe greater unmasking in the closed-ID task, less intense maskers in the open-ID may have reduced the overall unmasking independent of the task difficulty.

The results of this study provide insight into the effect of speech task set size on uncomodulated glimpsing. While overall unmasking may decline in the open-set task, it was possible that uncomodulated glimpsing would be increased in the open-set task relative to the closed-set task. An increase in uncomodulated glimpsing was hypothesized for the open-set protocol due to more importance placed on all available speech information. Since uncomodulated glimpsing was calculated as the difference between thresholds in the dichotic asynchronous AM conditions and those of the best dichotic control, insufficient cues in either the even or odd bands alone would impact the dichotic controls more severely in the open-set than the closed-set protocol. In fact, the results show that the least amount of glimpsing (in dB) in the open-set task was still greater than greatest amount of glimpsing in the closed-set task. This outcome supported our hypothesis that the open-set condition would be associated with reduced masking release and increased evidence of uncomodulated glimpsing.

Conclusions

The present study tested whether uncomodulated glimpsing in Howard-Jones and Rosen's (1993) study was limited by the peripheral phenomenon of USM, particularly for large numbers of bands. By presenting even bands and odd bands of asynchronous AM maskers to opposite ears, we have shown that significantly greater release from masking is possible with dichotic presentation even when maskers are filtered into as many as 16 bands. While no benefit to dichotic presentation over diotic presentations was observed in the open-

set task, it is likely that the low masker level at threshold played a role in this result. Future studies may show that at high enough presentation levels, the listeners are able to take advantage of dichotic presentations.

This study shows more evidence that normal hearing listeners are able to integrate speech information asynchronously masked in time and frequency. The current maskers are, however, predictable in their spectro-temporal structure, and therefore do not reflect the randomness of many natural masking environments. Nevertheless, this study has possible implications for hearing aid design for those with hearing impairment. For example, bilateral auditory prostheses could implement processing strategies that could ameliorate the disruptive effects of USM between neighboring frequency region by splitting even and odd numbered bands to opposite ears. Further studies would need to be conducted to show that hearing impaired listeners indeed benefit from dichotic listening of asynchronous AM maskers.

Figures

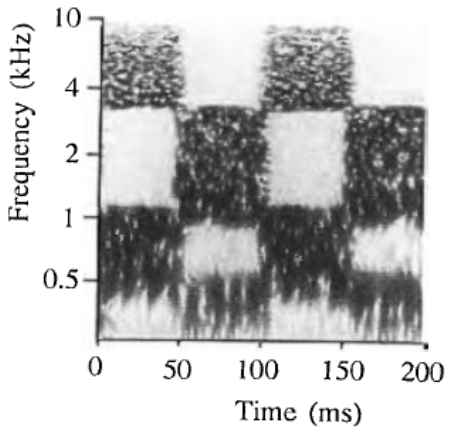

Figure 1: Spectrogram of 4-band checkerboard masker (Howard-Jones & Rosen, 1993).

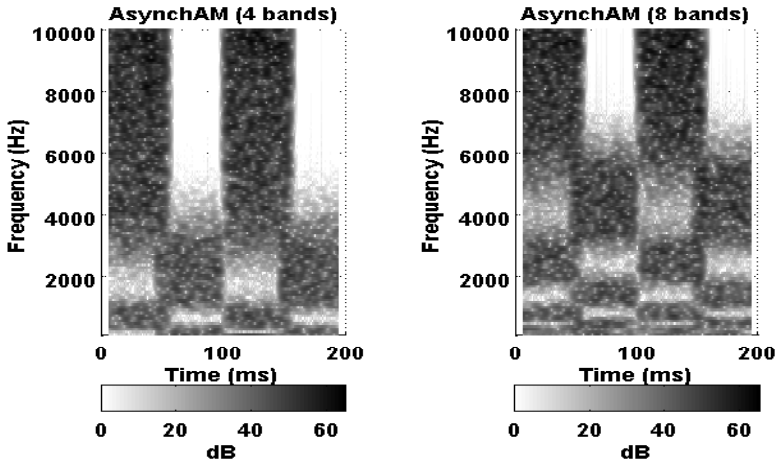

Figure 2: Time-frequency plots (spectrograms) of asynchronous noise at the output of a bank of auditory filters, simulated by 128 gammatone filters (Slaney, 1998). Higher energy in the noise is represented by dark shading, and regions with very low energy are represented by white. Regions of pink noise in the “off” phase of modulation are more greatly affected by upward spread of masking (USM) when the asynchronous AM noise is filtered into 8 bands (right side) versus 4 bands (left side).

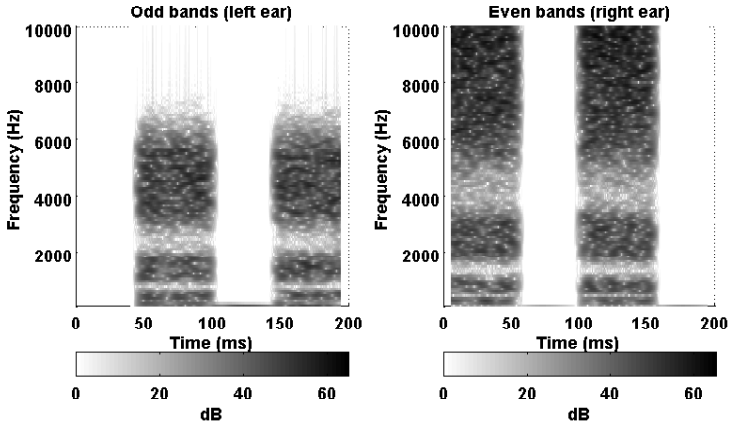

Figure 3: Time-frequency plots (spectrograms) of odd (left) and even (right) bands alone for an 8-band asynchronous AM masker at the output of a bank of auditory filters, simulated by 128 gammatone filters (Slaney, 1998). Contralateral upward spread of masking (USM) does not mask the ipsilateral signal.

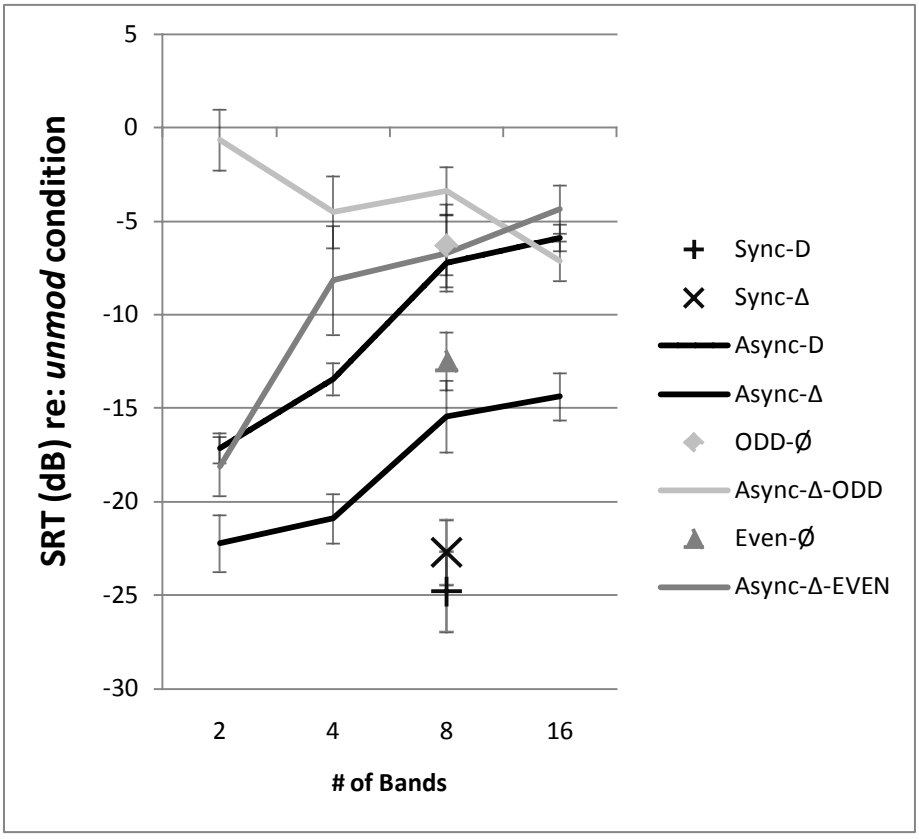

Figure 4: Speech reception thresholds (SRTs) in experiment 1a are plotted for amplitude modulated (AM) noise conditions relative to the control condition, Unmod. Error bars indicate standard error (n = 6).

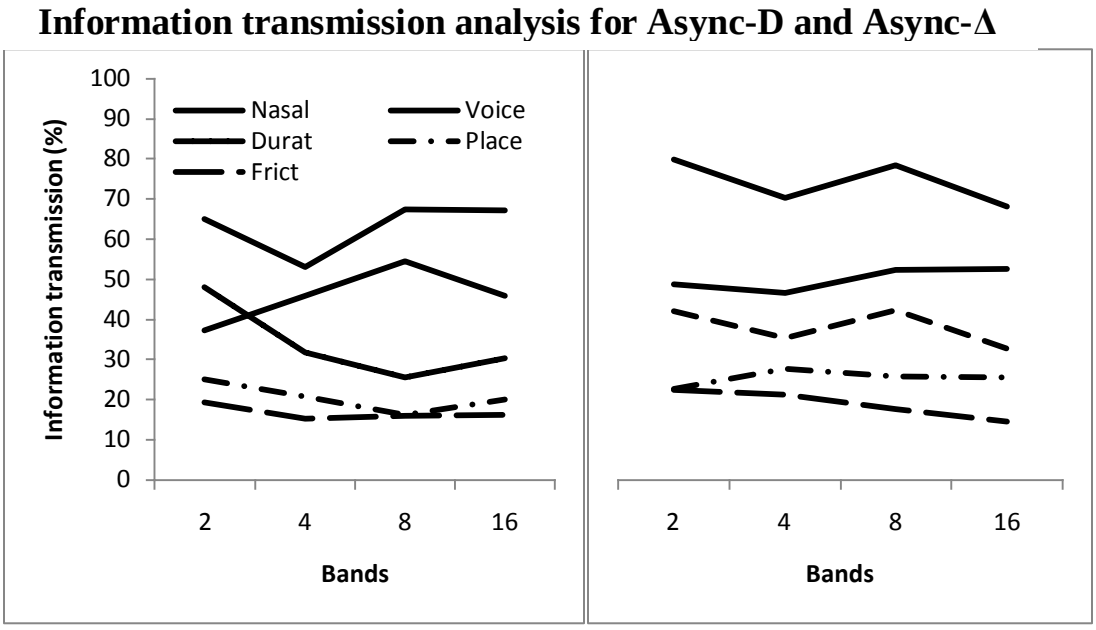

Figure 5: The relative information transmitted (in percent) is plotted as a function of number of bands for each feature in the diotic (left) and dichotic (right) asynchronous AM conditions of experiment 1a.

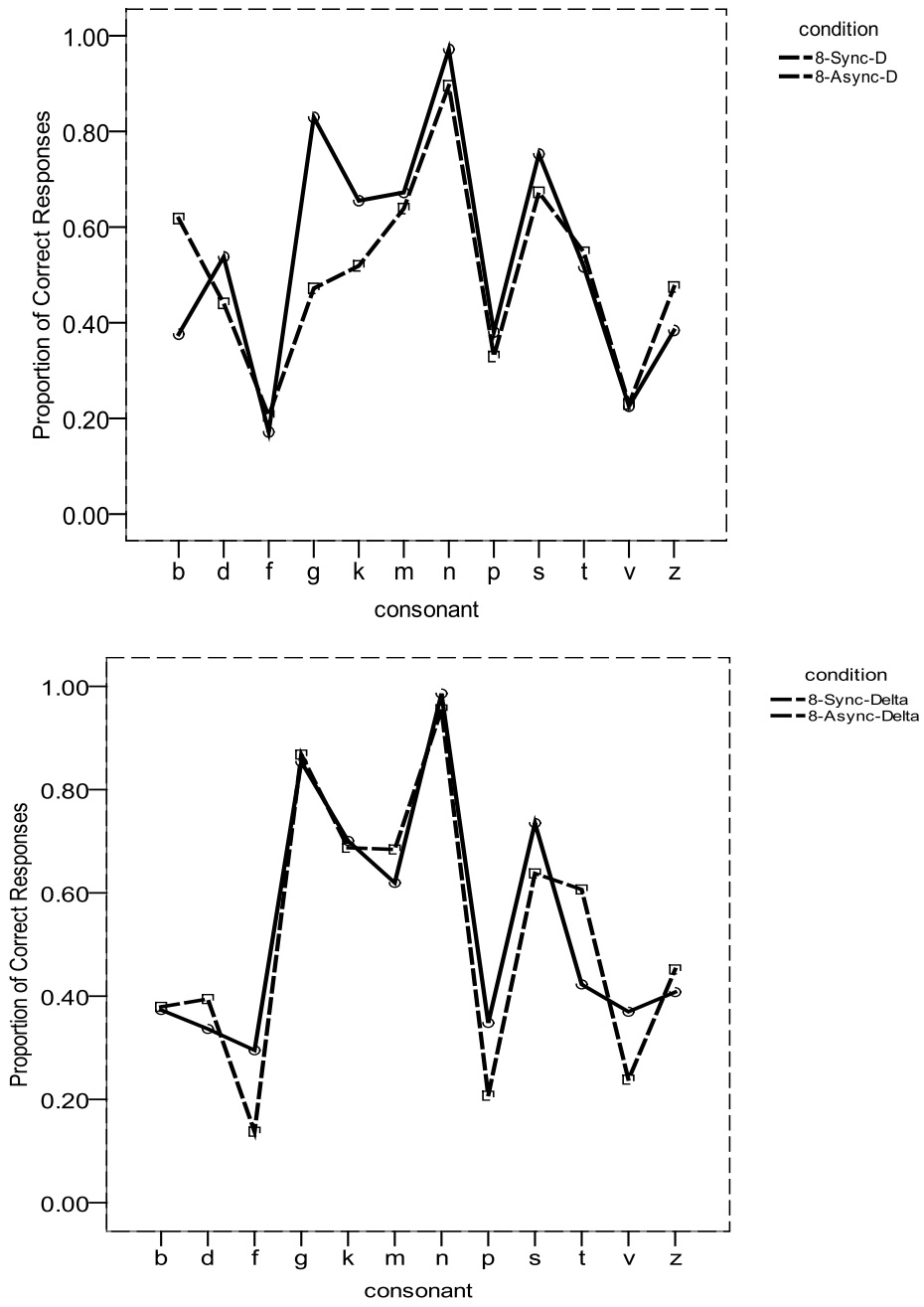

Figure 6: Average proportion of correct responses by consonant for 8-Sync and 8-Async conditions either presented diotically (top) or dichotically (bottom). Consonant accuracy varies widely, but is relatively independent of condition or stimulation type.

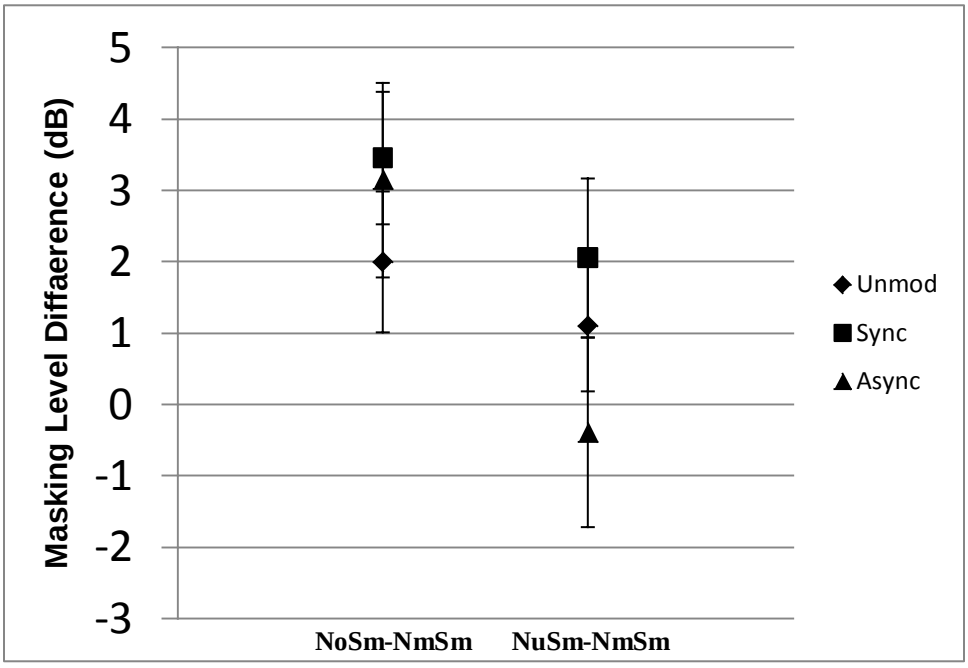

Figure 7: Masking-level differences (MLDs) from experiment 1b for three masking noise conditions (Unmod, Sync, and Async) are shown.

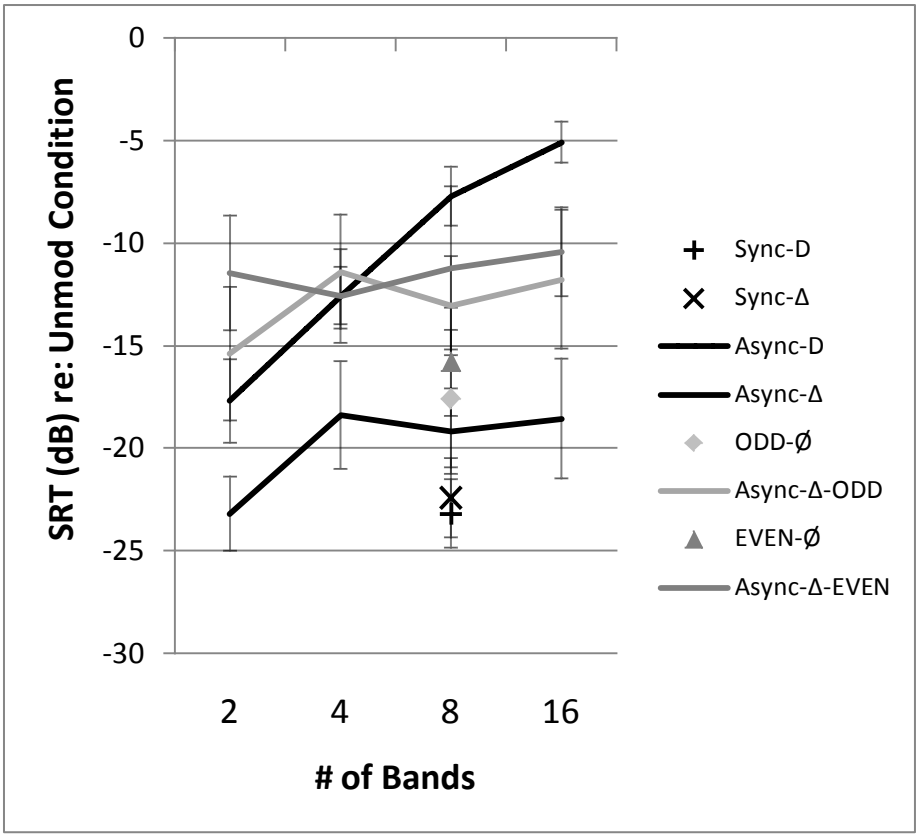

Figure 8: Speech reception thresholds (SRTs) for closed-set protocol in experiment 2 are plotted for amplitude modulated (AM) noise conditions relative to the control condition, Unmod. Error bars indicate standard error (n = 5).

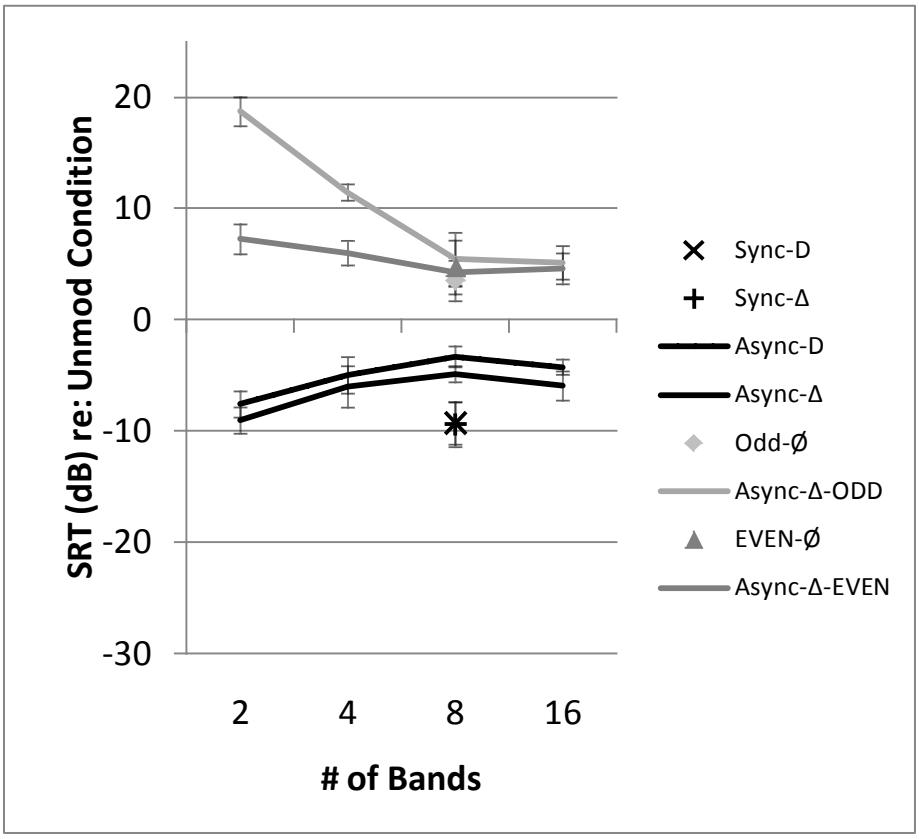

Figure 9: Speech reception thresholds (SRTs) for open-set protocol in experiment 2 are plotted for amplitude modulated (AM) noise conditions relative to the control condition, Unmod. Error bars indicate standard error (n = 5).

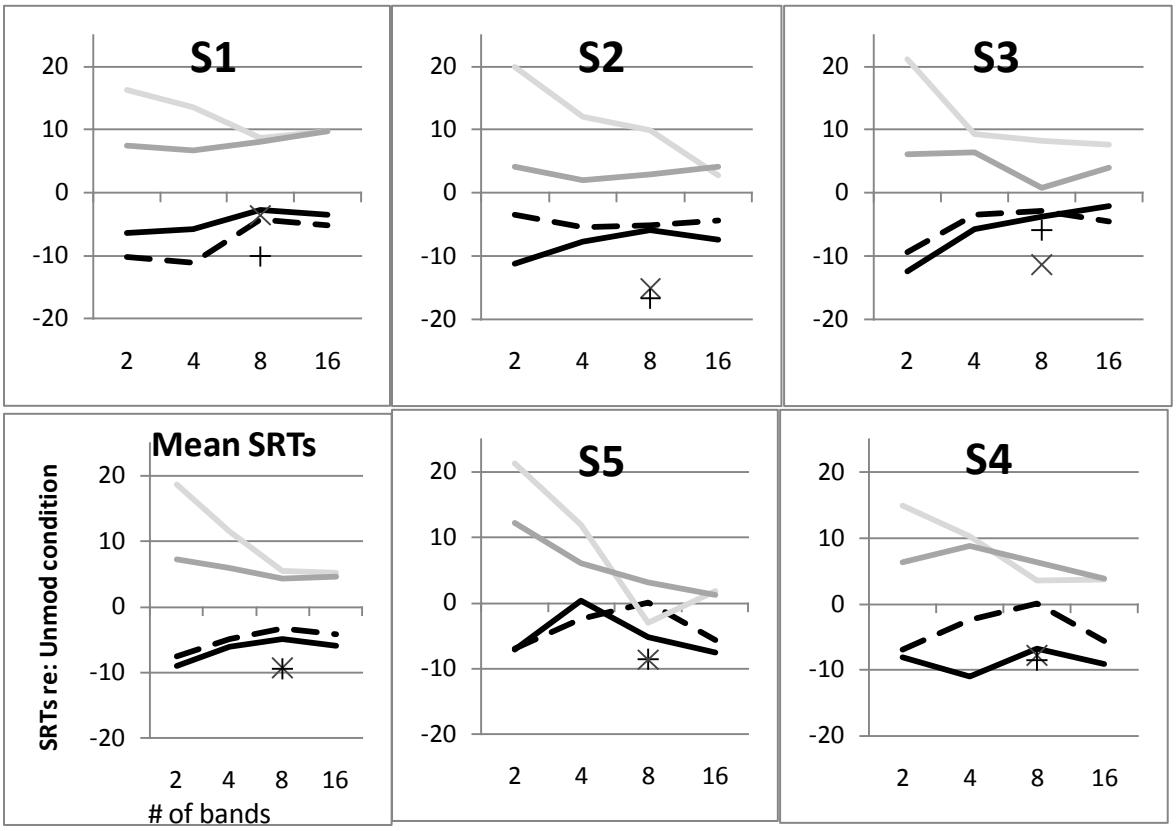

Figure 10: Individual SRTs in the open-set protocol for experiment 2 are plotted for AM noise conditions relative to the reference condition, Unmod. The Async-Δ-ODD and Async-Δ-EVEN (light grey and grey solid lines, respectively) are plotted with the Async-Δ and Async-D (solid and dotted black lines, respectively). Sync-D and Sync-Δ are also plotted (X marker and filled circle, respectively).

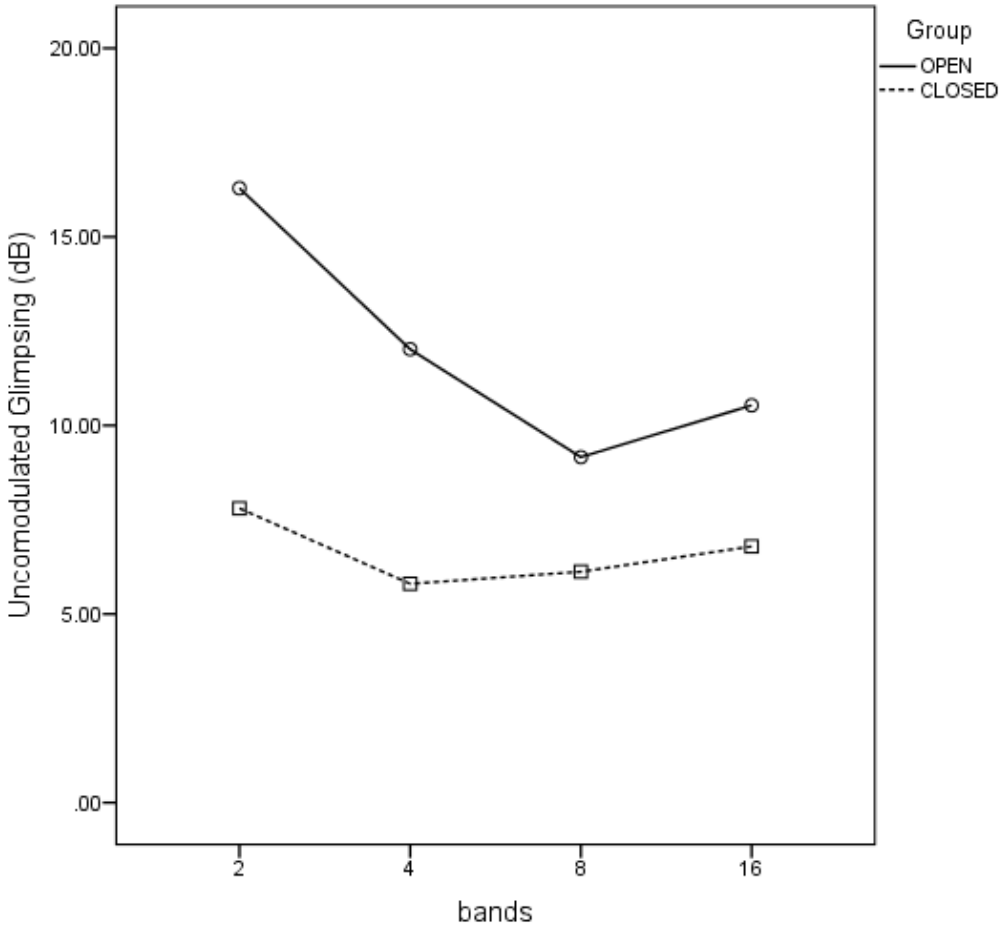

Figure 11: Uncomodulated glimpsing in experiment 2 calculated as the difference in SRT between the dichotic asynchronous condition and the best SRT of the two dichotic controls. Symbols and line styles indicate the test protocol, which was either open-set (circles, solid line) or closed-set (squares, dotted line). Although SRTs were generally worse in the open set protocol compared to the closed set, asynchronous glimpsing was found to be significantly greater in the open set.

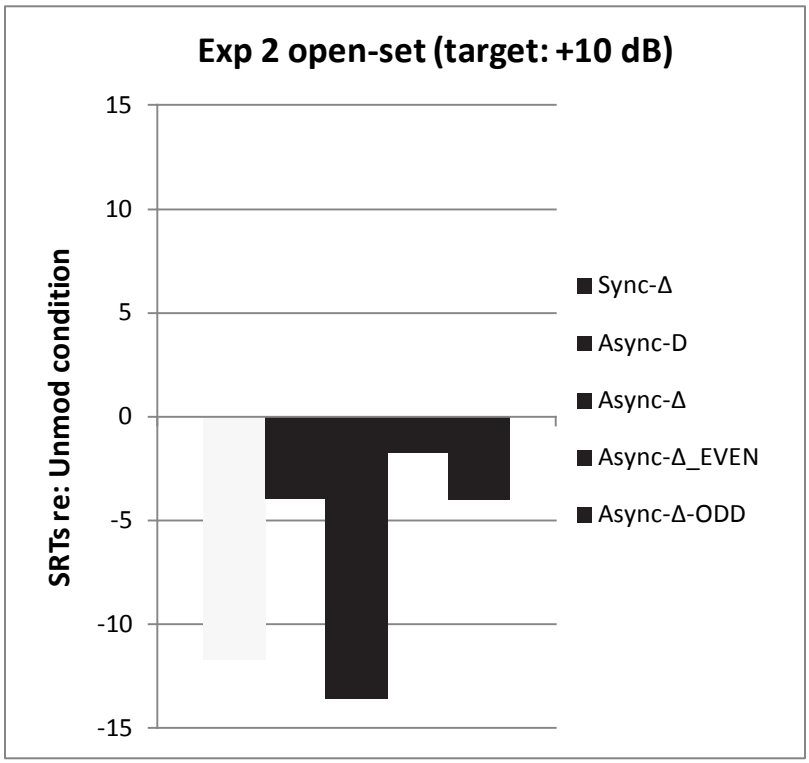

Figure 12: Pilot data for experiment 2 (open-set) in which the target level was increased by 10 dB. Data suggest that the advantage of dichotic presentation is evident for higher masker intensities.

Tables

Table 1: Conditions in experiments 1a and 2, and pilot data (experiment 3).

Condition	*Abbreviation*	Stimulation	Signal Presentation	Masker Modulation Pattern	Bands
1	Unmod	diotic	full	no AM	n/a
2	Sync-D	diotic	full	sync AM	n/a
3	2-Async-D	diotic	full	async AM	2
4	4-Async-D	diotic	full	async AM	4
5	8-Async-D	diotic	full	async AM	8
6	16-Async-D	diotic	full	async AM	16
7	8-ODD-ø	left ear	just-odd	just-odd	8
8	8-EVEN-ø	right ear	just-even	just-even	8
9	Sync-Δ	dichotic	split	sync AM	8
10	2-Async-Δ	dichotic	split	async AM	2
11	4-Async-Δ	dichotic	split	async AM	4
12	8-Async-Δ	dichotic	split	async AM	8
13	16-Async-Δ	dichotic	split	async AM	16
14	2-Async-Δ-ODD	dichotic	just-odd	async AM	2
15	4-Async-Δ-ODD	dichotic	just-odd	async AM	4
16	8-Async-Δ-ODD	dichotic	just-odd	async AM	8
17	16-Async-Δ-ODD	dichotic	just-odd	async AM	16
18	2-Async-Δ-EVEN	dichotic	just-even	async AM	2
19	4-Async-Δ-EVEN	dichotic	just-even	async AM	4
20	8-Async-Δ-EVEN	dichotic	just-even	async AM	8
21	16-Async-Δ-EVEN	dichotic	just-even	async AM	16

Table 2: Percent transmission (in bits per stimulus) for all conditions and for each feature separately (Miller and Nicely, 1955).[1]

Condition	*Abbreviation*	All	Voice	Nasal	Frict	Durat	Place
1	Unmod	42%	55	61	16	27	22
2	Sync-D	45	52	77	19	32	28
3	2-Async-D	38	37	65	19	48	25
4	4-Async-D	37	46	53	15	32	21
5	8-Async-D	40	54	67	16	26	16
6	16-Async-D	39	46	67	16	30	20
7	8-ODD-ø	44	58	62	15	22	32
8	8-EVEN-ø	44	54	86	22	40	19
9	Sync-Δ	45	54	74	16	33	31
10	2-Async-Δ	43	49	80	22	42	23
11	4-Async-Δ	41	47	70	21	35	28
12	8-Async-Δ	45	52	78	18	42	26
13	16-Async-Δ	44	53	68	14	33	26
14	2-Async-Δ-ODD	44	70	96	18	15	10
15	4-Async-Δ-ODD	44	56	86	11	8	30
16	8-Async-Δ-ODD	42	56	63	13	19	25
17	16-Async-Δ-ODD	40	52	65	20	37	19
18	2-Async-Δ-EVEN	39	33	57	21	52	32
19	4-Async-Δ-EVEN	44	60	68	28	62	14
20	8-Async-Δ-EVEN	44	55	86	26	48	17
21	16-Async-Δ-EVEN	42	51	61	27	56	26

[1] Importantly, the sum of the bits per channel should equal approximately the transmission calculated for the five channels taken together, but due to redundancy in speech, this sum is typically greater than the composite calculation.

Table 3: Confusion matrix for Sync-D condition in experiment 1a.

	/b/	/d/	/f/	/g/	/k/	/m/	/n/	/p/	/s/	/t/	/v/	/z/
/b/	**.40**	.06	.01	**.14**	.02	.09	.03	.03	.01		**.20**	.01
/d/		**.56**		**.30**			.02			.05		.07
/f/	**.15**		**.16**	.04	.07	.03	.01	.08	**.18**	**.16**	.08	.3
/g/		**.10**		**.81**	.03		.03			.01		
/k/			.05	.03	**.63**		.02	.2	.8	**.16**	.1	.1
/m/	.01			.01	.01	**.68**	**.28**					
/n/				.01	.01		**.98**					
/p/	.01		.07	.04	**.17**	.01		**.37**	.03	**.28**		.02
/s/		.01	.01	.03	.03				**.70**	.07		**.16**
/t/			.07	.03	**.23**		.01	.07	.08	**.51**		.01
/v/	**.23**	.06	.02	**.28**	.02	.02	.02		.01	.01	**.27**	.06
/z/	.04	**.15**		**.11**	.04			.01	**.14**	.04	.06	**.40**

Table 4: Confusion matrix for 8-Async-D condition in experiment 1a.

	/b/	/d/	/f/	/g/	/k/	/m/	/n/	/p/	/s/	/t/	/v/	/z/
/b/	**.68**	.01	.03	.01		.08	.01	.01			**.11**	.04
/d/	.05	**.41**		**.27**	.01	.01	.06		.02	.01	.01	**.13**
/f/	.01	.07	**.20**	.09	**.13**		.01	**.11**	**.16**	**.12**	.05	.05
/g/	.03	**.19**	.01	**.49**	.03	.01	**.12**	.01	.01	.03	.01	.04
/k/			.03	.04	**.51**	.01	.05	**.20**	.05	.09	.01	
/m/						**.66**	**.31**	.01		.01		
/n/						**.11**	**.88**				.02	
/p/			.08		**.24**	.01	.01	**.32**	.04	**.28**		
/s/	.03		.06	.04	.04		.01	.03	**.66**	.06	.01	.04
/t/		.01	.05		**.11**		.01	**.18**	.06	**.57**	.01	
/v/	**.33**	.07		**.15**	.08	.06	.04	.01			**.21**	.04
/z/	.06	**.11**		.04	.04		.04	.06	.06	.04	.08	**.47**

Table 5: Confusion matrix for 8-Async-Δ condition in experiment 1a.

	/b/	/d/	/f/	/g/	/k/	/m/	/n/	/p/	/s/	/t/	/v/	/z/
/b/	**.38**	.01	.01	**.18**	.03	**.10**	.09			.01	**.17**	.01
/d/	.02	**.38**		**.40**		.02	.06		.02		.02	.09
/f/	**.14**	.05	**.14**	.09	**.27**	.02	.01	.07	.01	.08	.08	.02
/g/		.03		**.88**	.03	.01	.01	.01		.01		.01
/k/		.01	.01	.04	**.70**		.02	.04		**.15**	.01	.02
/m/	.01					**.70**	**.26**		.01	.01	.01	
/n/						.03	**.96**				.01	
/p/	.02		.09	.03	**.37**			**.22**		**.27**		
/s/	.01		.06		.01	.01			**.62**			**.28**
/t/		.02	.02	.02	**.23**		.02	.03	.02	**.63**	.02	.02
/v/	**.21**	.05	.01	**.38**	.03		.05	.03		.01	**.23**	
/z/	**.11**	.08		**.24**			.01	.01	.04	.01	.04	**.46**

References

ANSI (**1996**). "ANSI S3.6-1996," in *American National Standards Specification for Audiometers* (American National Standards Institute, New York).

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

Bacon, S. P., Opie, J. M., and Montoya, D. Y. (**1998**). "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," J. Speech. Lang. Hear. Res. **41**, 549-563.

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acustica **86**, 117-128.

Brungart, D. S., and Simpson, B. D. (**2002**). "Within-ear and across-ear interference in a cocktail-party listening task," J. Acoust. Soc. Am. **112**, 2985-2995.

Brungart, D. S., Simpson, B. D., Darwin, C. J., Arbogast, T. L., and G. Kidd, J. (**2005**). "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **117**, 292-304.

Buss, E., Hall, J. W., 3rd, and Grose, J. H. (**2003**). "Effect of amplitude modulation coherence for masked speech signals filtered into narrow bands," J. Acoust. Soc. Am. **113**, 462-467.

Buss, E., Hall, J. W., 3rd, and Grose, J. H. (**2004**). "Spectral integration of synchronous and asynchronous cues to consonant identification," J. Acoust. Soc. Am. **115**, 2278-2285.

Buss, E., Whittle, L. N., Grose, J. H., and Hall, J. W., 3rd (**2009**). "Masking release for words in amplitude-modulated noise as a function of modulation rate and task," J. Acoust. Soc. Am. **126**, 269-280.

Buus, S. (**1985**). "Release from masking caused by envelope fluctuations," J. Acoust. Soc. Am. **78**, 1958-1965.

Carhart, R., Tillman, T. W., and Johnson, K. R. (**1966**). "Binaural masking of speech by periodically modulated noise," J. Acoust. Soc. Am. **39**, 1037-1050.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975-979.

Culling, J. F., and Colburn, H. S. (**2000**). "Binaural sluggishness in the perception of tone sequences and speech in noise," J. Acoust. Soc. Am. **107**, 517-527.

Dirks, D. D., and Bower, D. (**1970**). "Effect of forward and backward masking on speech intelligibility," J. Acoust. Soc. Am. **47**, 1003-1008.

Dirks, D. D., and Wilson, R. H. (**1969**). "The effect of spatially separated sound sources on speech intelligibility," J. Speech Hear. Res. **12**, 5-38.

Edmonds, B. A., and Culling, J. F. (**2006**). "The spatial unmasking of speech: evidence for better-ear listening," J. Acoust. Soc. Am. **120**, 1539-1545.

Eisenberg, L. S., Dirks, D. D., and Bell, T. S. (**1995**). "Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing," J. Speech Hear. Res. **38**, 222-233.

Festen, J. M., 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. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578-3588.

George, E. L., Festen, J. M., and Houtgast, T. (**2006**). "Factors affecting masking release for speech in modulated noise for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, 2295-2311.

Gnansia, D., Jourdes, V., and Lorenzi, C. (**2008**). "Effect of masker modulation depth on speech masking release," Hear. Res. **239**, 60-68.

Grose, J. H., and Hall, J. W., 3rd (**1992**). "Comodulation masking release for speech stimuli," J. Acoust. Soc. Am. **91**, 1042-1050.

Hall, J. W., Grose, J. H., and Dev, M. B. (**1997**). "Signal detection and pitch ranking in conditions of masking release," J. Acoust. Soc. Am. **102**, 1746-1754.

Hall, J. W., Haggard, M. P., and Fernandes, M. A. (**1984**). "Detection in noise by spectro-temporal pattern analysis," J. Acoust. Soc. Am. **76**, 50-56.

Hirsh, I. J. (**1948**). "Binaural summation and interaural inhibition as a function of the level of masking noise," Am. J. Psychol. **61**, 205-213.

Hirsh, I. J. (**1950**). "The relation between localization and intelligibility," J. Acoust. Soc. Am. **22**, 196-200.

Howard-Jones, P. A., and Rosen, S. (**1993**). "Uncomodulated glimpsing in "checkerboard" noise," J. Acoust. Soc. Am. **93**, 2915-2922.

Kwon, B. J. (**2002**). "Comodulation masking release in consonant recognition," J. Acoust. Soc. Am. **112**, 634-641.

Levitt, H., and Rabiner, L. R. (**1967a**). "Binaural release from masking for speech and gain in intelligibility," J. Acoust. Soc. Am. **42**, 601-608.

Levitt, H., and Rabiner, L. R. (**1967b**). "Predicting binaural gain in intelligibility and release from masking for speech," J. Acoust. Soc. Am. **42**, 820-829.

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

Martin, F. N., Bailey, H. A. T., and Pappas, J. J. (**1965**). "The effect of central masking on threshold for speech," J. Aud. Res. **5**, 293-296.

Martin, F. N., and Digiovanni, D. (**1979**). "Central masking effects on spondee threshold as a function of masker sensation level and masker sound pressure level," J. Am. Aud. Soc. **4**, 141-146.

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

Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338-352.

Moore, B. C., Alcantara, J. I., and Dau, T. (**1998**). "Masking patterns for sinusoidal and narrow-band noise maskers," J. Acoust. Soc. Am. **104**, 1023-1038.

Moore, B. C. J. (**2003**). *An introduction to the psychology of hearing* (Academic Press, Amsterdam ; Boston).

Nelken, I., Rotman, Y., and Bar Yosef, O. (**1999**). "Responses of auditory-cortex neurons to structural features of natural sounds," Nature **397**, 154-157.

Peters, R. W., Moore, B. C., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577-587.

Saberi, K., Dostal, L., Sadralodabai, T., Bull, V., and Perrott, D. R. (**1991**). "Free-field release from masking," J. Acoust. Soc. Am. **90**, 1355-1370.

Slaney, M. (**1998**). "Auditory Toolbox Version 2," in *Technical Report #1998-010* (Interval Research Corporation).

Smith, D. W., Turner, D. A., and Henson, M. M. (**2000**). "Psychophysical correlates of contralateral efferent suppression. I. The role of the medial olivocochlear system in "central masking" in nonhuman primates," J. Acoust. Soc. Am. **107**, 933-941.

Summers, V., and Molis, M. R. (**2004**). "Speech recognition in fluctuating and continuous maskers: effects of hearing loss and presentation level," J. Speech Lang. Hear. Res **47**, 245-256.

Wegel, R. L., and Lane, C. E. (**1924**). "The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear," Physical Review **23**, 266.

Wilson, R. H., and Carhart, R. (**1969**). "Influence of pulsed masking on the threshold for spondees," J. Acoust. Soc. Am. **46**, 998-1010.

Wilson, R. H., Hopkins, J. L., Mance, C. M., and Novak, R. E. (**1982**). "Detection and recognition masking-level differences for the individual CID W-1 spondaic words," J. Speech Hear. Res. **25**, 235-242.

Zwislocki, J. J. (**1971**). "Central masking and neural activity in the cochlear nucleus," Audiology **10**, 48-59.