Perception & Psychophysics
1992, 52 (4), 437-445

The effect of amplitude comodulation on auditory object formation in sentence perception

THOMAS D. CARRELL and JANE M. OPIE
Northwestern University, Evanston, Illinois

To comprehend speech in most environments, listeners must combine some but not all sounds from across a wide range of frequencies. Three experiments were conducted to examine the role of amplitude comodulation in performing an essential part of this function: the grouping together of the simultaneous components of a speech signal. Each of the experiments used time-varying sinusoidal (TVS) sentences (Remez, Rubin, Pisoni, & Carrell, 1981) as base stimuli because their component tones are acoustically unrelated. The independence of the three tones reduced the number of confounding grouping cues available compared with those found in natural or computer-synthesized speech (e.g., fundamental frequency and simultaneity of harmonic onset). In each of the experiments, the TVS base stimuli were amplitude modulated to determine whether this modulation would lead to appropriate grouping of the three tones as reflected by sentence intelligibility. Experiment 1 demonstrated that amplitude comodulation at 100 Hz did improve the intelligibility of TVS sentences. Experiment 2 showed that the component tones of a TVS sentence must be comodulated (as opposed to independently modulated) for improvements in intelligibility to be found. Experiment 3 showed that the comodulation rates that led to intelligibility improvements were consistent with the effective rates found in experiments that examined the grouping of complex nonspeech sounds by common temporal envelopes (e.g., comodulation masking release; Hall, Haggard, & Fernandes, 1984). The results of these experiments support the claim that certain basic temporal-envelope processing capabilities of the human auditory system contribute to the perception of fluent speech.

Natural environments typically present the listener with sounds from many sources simultaneously, and the auditory perceptual system must parse complex incoming waveforms into their original sources for further processing. Most speech-perception research over the past several decades has either ignored this problem or proceeded under the assumption that the parsing has already occurred. Major findings in speech perception have been derived from experiments conducted in artificially quiet environments and with stimuli presented via headphones. The reason for this was simple. Speech itself is a very complex signal that is difficult to study. When combined with the fact that the transmission channels provided by natural environments are also extremely complex, the investigation of speech perception had to be simplified by ignoring the effect of competing sounds.

Recent work, however, has addressed many issues relevant to the study of speech in natural environments. For example, there is now evidence that some of the structure in the speech signal that was previously ignored because it was linguistically irrelevant or difficult to take into account may be of significant perceptual value in separating the auditory foreground from its background. The process of "auditory object formation" (Moore, 1989), also known as auditory image formation (McAdams, 1984), now appears relevant to understanding speech perception in natural listening environments. Although this concept was originally devised to describe how a listener separates the auditory figure from the ground in music and general auditory perception, there is evidence that auditory object formation is valuable in the process of speech perception as well.

A number of processes in auditory perception determine which particular sounds of a complex signal are grouped together to be heard as a single auditory object (see Bregman, 1990). Some of these processes have been shown to cause the components of a speech signal to cohere as a unit so that, for example, the formants of a syllable are perceived as one unit rather than as separate acoustic events. Two studies, both by Darwin, illustrate examples of auditory object formation in speech perception. In the first study, fundamental frequencies of the formants of a syllable were shown to control which formants were grouped together and which were excluded from listeners' phonetic percepts (Darwin, 1981). This was demonstrated with unusual four-formant speechlike syllables, which were perceived as /ru/if the fourth formant was excited by a different fundamental frequency than the

Preparation of this article was supported in part by Northwestern University URGC Awards 03XE and 05XF to T.D.C. Portions of this work were presented to the 116th and 118th meetings of the Acoustical Society of America. We would like to thank the anonymous reviewers for their valuable comments. Address correspondence to T. D. Carrell, Northwestern University, Department of Communication Sciences and Disorders, 2299 Sheridan Road, Evanston, IL 60208-3570 (e-mail: tcarrell@nwu.edu). Jane Opie is now at the Department of Speech and Hearing Sciences, Arizona State University, Tempe, AZ 85287-0102.

other formants but as /li/ if the second formant was excited by the odd fundamental. In a second study, the harmonic fine structure of syllables was shown to influence auditory object formation (Darwin, 1984). When the onset and offset of the harmonics of an isolated vowel were experimentally manipulated, it was found that harmonics that started and stopped simultaneously contributed to a vowel's identity, whereas individual harmonics that were slightly offset from the rest did not. These two studies illustrate that there is nonphonetic structure in the acoustic waveform that can group appropriate sounds together and exclude others. Furthermore, they indicate that there is reason to examine the entire speech signal for additional grouping cues before relying on such well-accepted effects as context and listener expectation to explain the ability of humans to extract speech from noisy, natural environments.

Although the segregation of speech from background noise was not directly addressed in this study, a crucial step in this process, the grouping of simultaneous sounds together into an auditory object, was. The specific goal of the experiments that follow was to demonstrate that amplitude modulation at the fundamental frequency provides strong grouping capabilities in sentence perception, which then leads to improved intelligibility. There are two major factors that lead one to expect this finding. The first factor is based on the way sounds are created in the natural environment. Speech sounds have much of their energy generated by vibration of the vocal folds. This creates a natural periodic variation in amplitude at the fundamental frequency.[1] In addition, many nonspeech sounds consist of a fundamental frequency with harmonic energy, often creating a waveform envelope with natural amplitude modulation. Therefore, from a straightforward problem-solving perspective, a good strategy for grouping a complex array of simultaneous sounds into their sources would be to associate those spectral components with common amplitude modulation (or comodulation) together and to exclude other sounds. In fact, strong evidence for this sort of grouping has already been found with nonspeech stimuli (Bregman, Abramson, Doehring, & Darwin, 1985).

The second reason that amplitude modulation is likely to be important in auditory object formation is based on the phenomenon of comodulation masking release (CMR) (Hall, Haggard, & Fernandes, 1984; McFadden, 1987; Schooneveldt & Moore, 1987) and related temporal auditory processes such as modulation detection interference (Yost, Sheft, & Opie, 1989) and comodulation difference detection (Cohen & Schubert, 1987; McFadden, 1987). The basic finding is that a tone presented at subthreshold levels and centered in an amplitude-modulated narrowband noise may become audible if another noise band is *added* at a different frequency. That is, more noise leads to better signal detection. This only occurs, however, if the second noise band is amplitude modulated at the same rate and phase as the first. Therefore, adding more noise to a tone already presented in noise can make the tone more audible, but only if the noises are properly comodulated. One interpretation of the CMR effect is that the noise bands in the preceding example were grouped together by common amplitude modulation and this grouping made the excluded tone more perceptible (Yost & Sheft, 1989). The CMR literature provides evidence that the grouping of amplitude-comodulated components of a frequency spectrum is a basic capability of the auditory system (e.g., Hall, 1987). Therefore, one might expect that speech-perception processes would take advantage of this general perceptual capability.[2]

Despite the evidence that amplitude comodulation can form auditory objects in relatively simple stimuli, and in addition to the rationale for its usefulness in separating natural sound sources, there has been no direct evidence that comodulation is useful in the perception of sentence-length utterances. The following experiments extend our knowledge of the effect of amplitude comodulation on grouping from the domain of general auditory perception to the domain of speech perception. This grouping is assumed to be an important step in separating a speech signal from its background.

EXPERIMENT 1

Time-varying sinusoidal (TVS) sentences (Remez, Rubin, Pisoni, & Carrell, 1981) were employed to test amplitude comodulation as a grouping cue in speech perception. These stimuli were chosen because they are free from most known grouping cues, allowing simpler interpretation of the grouping effects found with further stimulus manipulations. The TVS sentences in the experiments reported here were constructed with only three tones each of a constant amplitude that followed the center frequencies of the formants of naturally spoken sentences. They had no fundamental frequency, no harmonic structure, and no internal amplitude onsets or offsets.

The TVS sentences were then amplitude modulated at 100 Hz, creating a second set of stimuli. The gross effect of amplitude modulating a TVS signal is to rapidly turn on and off each of the three component tones simultaneously, that is, to comodulate the three tones. An increase in the intelligibility of amplitude-modulated TVS sentences as compared with unmodulated TVS sentences would support the idea that amplitude comodulation grouped the three tones together into an auditory object. This is because such a grouping should make the signal easier to process phonetically than if it were necessary to combine three independent tones at a higher (e.g., lexical or semantic) level.

Method

Subjects. Twenty-eight subjects were recruited from the students and staff of the Department of Communication Sciences and Disorders at Northwestern University. They had a mean age of 24.1 years, with a range of 22-29 years. The subjects were paid with a $2.50 gift certificate for Mrs. Field's Cookies for their participation in this 15-min experiment. All subjects reported no current

or past speech or hearing problems, and none had heard TVS or TVS-based sentences prior to this experiment.

Stimuli. The experimental session was divided into two phases: identification and naturalness. A different set of stimuli was used in each phase.

Two types of stimuli were used in the identification phase: TVS sentences and amplitude-modulated TVS (AMTVS) sentences. The sentences were based on five naturally produced utterances: "Hello Lenny, how are you?"; "Where were you a year ago?"; "A yellow lion roared."; "We owe you a yo-yo."; and "When were you well?". These sentences will be referred to as LENNY, WHERE, YELLOW, YO-YO, and WHEN, respectively. The formant center frequencies were extracted at Haskins Laboratories (using the procedure described by Remez & Rubin, 1990) and were supplied to the author as data files containing the frequencies and amplitudes of the sentences at 10-msec intervals. The frequencies were entered into TONE (Kewley-Port, 1976), a synthesis program that constructed waveforms made up of independent sine waves. In the present experiments, the amplitudes were fixed at 60, 56, and 50 dB for the tones corresponding to Formants 1, 2, and 3, respectively. Note that the selected sentences were composed almost entirely of sonorants. TVS sentences containing stops and fricatives were not employed in this experiment because proper synthesis would have required amplitude variation in the tones over the course of the sentences (at the very least, in a binary fashion for stops). Sonorant-only sentences were chosen so that the amplitude of the tones could be kept constant throughout the stimulus. Any variation in amplitude would have provided another, confounding, auditory grouping cue. A narrowband spectrogram of the TVS sentence YELLOW produced in this manner is shown in Figure 1.

The AMTVS sentences were created by amplitude modulating each of the TVS sentences at 100 Hz. The modulating signal was a triangular wave with a duty cycle of 70%. On each 10-msec modulating cycle, the amplitude rose to 100% in 5 msec, dropped to 0% in 2 msec, and remained at 0% for the final 3 msec. The AMTVS stimulus construction technique is illustrated in Figure 2, and a narrowband spectrogram of the resulting signal is shown in Figure 3.

Note that the spectrograph of the AMTVS sentence is somewhat closer in appearance to a naturally produced sentence than is the TVS sentence. This is due to the 100-Hz sidebands flanking each of the three center frequencies. These sidebands should not be confused with harmonics; they are not harmonically related by an underlying fundamental frequency or to the sidebands of the other two tones.

For the naturalness phase of the experiment, two different types of stimuli were added to those used in the identification phase: natural sentences and amplitude-modulated natural sentences. The natural sentences consisted of four sentences taken from the Harvard Psychoacoustic Sentence List (Egan, 1948) spoken by a male native speaker of English. The amplitude-modulated natural sentences were created by the same modulating waveform as the AMTVS sentences. The purpose of including the sentences based on natural speech was to present listeners with a wide range of qualities for the naturalness rating task.

All sentences were constructed using a sampling rate of 10000 Hz and were presented under computer control with a 12-bit digital-to-analog converter. The sentences were then low-pass filtered at 4200 Hz with a Krohn-Hite 3343 filter configured for 96 dB per octave attenuation outside the passband. They were presented to

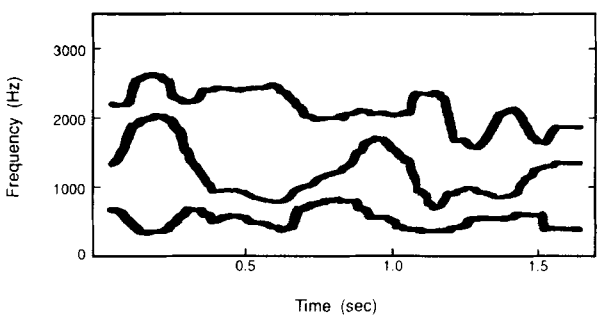

Figure 1. Narrowband spectrogram of the time-varying sinusoidal sentence: "A yellow lion roared."

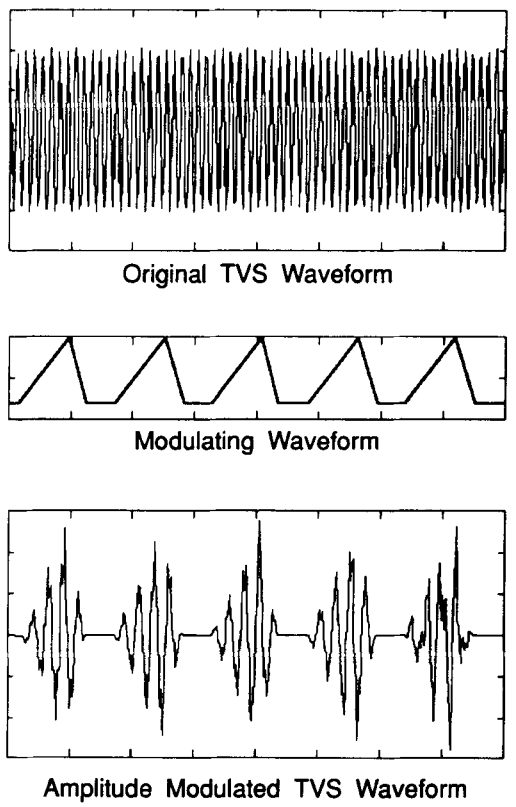

Figure 2. Construction technique used in the creation of amplitude-modulated time-varying sinusoidal (TVS) sentences. The top panel shows the time-domain waveform of a short portion of a TVS sentence. The center panel shows the triangular waveform used to modulate the original TVS waveform. The bottom panel shows the resulting amplitude-modulated waveform.

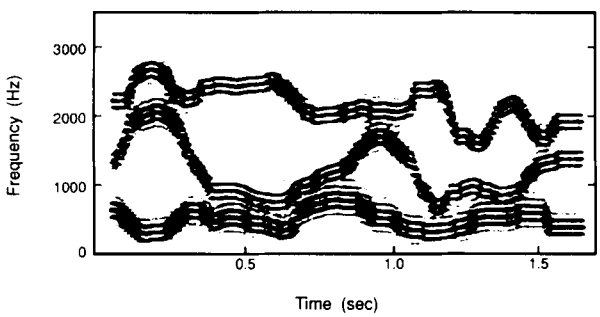

Figure 3. Narrowband spectrogram of the amplitude-modulated time-varying sinusoidal sentence: "A yellow lion roared."

the listeners binaurally at a peak level of 72 dB SPL with Sennheiser HD430 headphones.

Procedure. The listeners were first familiarized with the types of sounds they would hear. Familiarization consisted of instructing the subjects that they would be hearing the sample sentence "When were you well?", which was presented over headphones in both TVS and AMTVS formats. Each form of the sentence was repeated three times. Familiarization was conducted because pilot experiments had shown that many listeners required several presentations to hear TVS sentences as speech. Presenting the sentence types before collecting data was intended to reduce response variability caused by the listeners' gaining the ability to hear TVS sentences as speech at uncontrolled points during the testing trials.

In the identification phase of the experiment, two groups of subjects listened to four sentences each. Group 1 heard the first two sentences in TVS format and the second two in AMTVS format. Group 2 heard the first two sentences in AMTVS format and the second two in TVS format. This counterbalancing was performed to eliminate differences due to intrinsic sentence intelligibility. Each of the four sentences was presented three times with a 4-sec interstimulus interval (ISI). After the third presentation of each sentence, the subjects were given a 30-sec response interval to write the sentence they heard on their response forms. Beforehand, they were told that the sentences might be difficult and that they should make their best guesses if they were not certain about a sentence.

In the naturalness rating phase of the experiment (which immediately followed the intelligibility phase), 16 sentences were presented in random order. These were two versions (TVS and AMTVS) of the four sentences used in the identification phase and two versions (natural and AM natural) of four other sentences. The natural sentences were included to increase the range of the naturalness responses and were not subject to later analyses. The listeners were presented with each sentence twice. After the second presentation, they were instructed to select a number from 1 to 5 to rate the naturalness of each sentence. The ISI was 4 sec, and the response interval was 10 sec. The subjects were instructed to circle 1 if the sentence was "very unnatural, machine, or animal sounding," to circle 5 if the sentence was "very natural and human-like," and to use the values between for intermediate levels of naturalness. Sentence order was counterbalanced between the two groups.

Results

The first experiment demonstrated that the AMTVS sentences were both more intelligible and more natural than the simple TVS sentences.

To measure intelligibility, the subjects' orthographic responses were first converted to phonemes and were then scored on the basis of number of phonemes correct. These scores showed that AMTVS sentences were identified about 32% more accurately than were TVS sentences. The mean phonetic accuracies, broken down by sentence and modulation type, are shown in Figure 4.

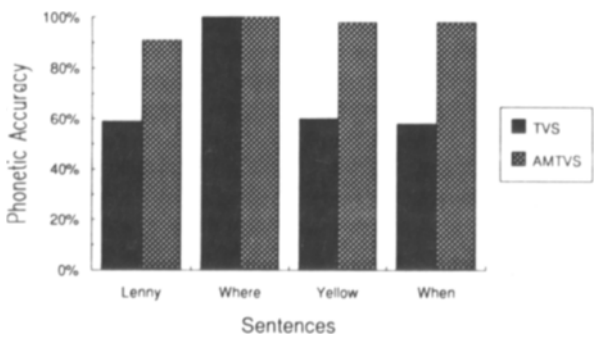

Figure 4. The intelligibility of modulated versus unmodulated time-varying sinusoidal sentences from Experiment 1.

Table 1
Mean Naturalness Ratings for TVS and AMTVS Sentences in Experiment 1

Modulation Type	Sentences			
	Lion	Where	Lenny	When
No modulation (TVS)	1.32	1.82	1.46	1.28
Amplitude modulation (AMTVS)	2.04	2.20	2.09	1.93

An independent t test was used to compare the phonetic accuracy of TVS versus AMTVS versions of each sentence. Three of the four sentences showed a significant difference between groups [$t(26) = 3.14, p < .005$, for LENNY; $t(26) = 4.90, p < .0001$, for YELLOW; and $t(26) = 4.25, p < .0002$, for YO-YO]. The sentence WHERE showed no significant difference between the TVS and the AMTVS conditions [$t(26) = 1.58$]. Inspection of the means revealed a ceiling effect that may have accounted for the lack of significance.

The results from the naturalness phase were examined with a 2 × 2 × 4 (sentence order × modulation × sentence) analysis of variance. The main effect of modulation [$F(1,26) = 106, p < .0001$] showed that AMTVS sentences were rated more natural than simple TVS sentences. There was also a main effect of sentence [$F(3,78) = 4.017, p < .01$]. There was no significant effect of sentence order or interactions. The means of the naturalness ratings are presented in Table 1.

Discussion

At least two broad classes of explanation exist for the intelligibility increment found with comodulated waveforms. One relies on auditory object formation and is based on physical properties of the stimulus. Another relies on the assumptions that amplitude modulation creates a more natural sounding sentence and that increased naturalness leads directly to increased intelligibility.

The auditory object explanation depends on the notion that it is difficult for a listener to group together the three acoustically unrelated tones composing a TVS sentence but easy to group together the simultaneous components of an AMTVS sentence. Because the three tones represent a single speech signal, any characteristic that causes them to be grouped as a unit should aid speech perception. Because of the effects of comodulation, AMTVS sentences have an acoustic structure more conducive to grouping than do TVS sentences. The observed intelligibility increment for the AMTVS sentences supports the auditory-object explanation.

A naturalness-based explanation could be also supported by the results of the rating phase of the first experiment. AMTVS sentences were judged to be more natural than simple TVS sentences. Therefore, it might be argued that increased naturalness would directly lead to increased intelligibility by causing the signal to be processed by speech-specific perceptual mechanisms. There are sev-

eral physical characteristics of an AMTVS sentence that make it more natural sounding than a TVS sentence. First, the sidebands created by amplitude modulation give the tones a bandwidth more similar to natural speech sounds. Second, the modulation may also supply a pitch (presumably not based on a greatest-common-divisor mechanism; Goldstein, 1973) that the listener may assume is the fundamental frequency. These speechlike characteristics of the AMTVS signal might prompt the perceptual system to process the input as a speech, rather than a nonspeech, signal. The next experiment was designed to dissociate naturalness from intelligibility and thereby contrast the two explanations outlined above.

EXPERIMENT 2

A different speechlike signal was developed to test the auditory-object explanation versus the naturalness explanation for the results found in the first experiment. As in Experiment 1, the new stimuli used as a base TVS sentences, which were then amplitude modulated. However, in this case, the amplitude modulation was performed separately for each of the three component tones in a sentence and each tone was modulated at a different rate. Because of the strategy of separate modulation, there was no comodulation of tones. These stimuli were named conflicting AMTVS (CAMTVS) sentences and were analogous to the uncorrelated stimuli used in many nonspeech experiments on CMR (e.g., Wright & McFadden, 1990). These signals preserved the spectral shape similarities to speech that are found in AMTVS sentences and had a sound quality similar to AMTVS sentences. It was expected that the CAMTVS sentences would be judged to be as natural as the AMTVS sentences, and having naturalness held constant would make it much easier to interpret the intelligibility findings. It was further expected that the CAMTVS sentences would be less intelligible than the AMTVS sentences because the listeners would not have amplitude comodulation available as a grouping cue for the three tones. These results would support the idea that amplitude comodulation serves a function in the formation of auditory objects.

Method

The method was the same as that used in Experiment 1 with the following exceptions.

Subjects. Thirty-three subjects were recruited from students in the Department of Communication Sciences and Disorders at Northwestern University. Their ages ranged from 18 to 34 years, with a mean age of 20.7. They were each paid $5 for their participation in this 30-min experiment.

Stimuli. Two types of stimuli were constructed for the identification phase of Experiment 2: AMTVS sentences and CAMTVS sentences. The AMTVS sentences were identical to the ones described in Experiment 1. The CAMTVS sentences were constructed by synthesizing each time-varying tone separately using the program TONE (Kewley-Port, 1976). Each tone was then amplitude modulated at a different frequency: Tone 1 at 97 Hz, Tone 2 at 79 Hz, and Tone 3 at 113 Hz. Care was taken to choose prime numbers for these frequencies so that inadvertent comodulation would not occur at the lowest common denominator of the three frequencies chosen. The resulting modulated tones were then digitally mixed so that the amplitudes of Tones 1, 2, and 3 were 60, 56, and 50 dB, respectively. Figure 5 shows a narrowband spectrogram of a CAMTVS sentence. Note that the CAMTVS and AMTVS sentences were both more similar in spectral appearance to natural speech than were the simple TVS sentences.

The naturalness phase of Experiment 2 contained three TVS, three AMTVS, and three CAMTVS sentences. There were no natural-sentence tokens in this set, so naturalness ratings cannot be directly compared with those measured in Experiment 1.

Procedure. Experiment 2 was designed much like Experiment 1. The listeners were first familiarized with the sound of AMTVS and CAMTVS sentences. Specifically, they were told that they would hear "A yellow lion roared." This sentence was then presented three times in both the AMTVS and CAMTVS format. After familiarization, the subjects were given an identification test, followed by a naturalness rating task.

In the identification phase of the experiment, two groups of subjects listened to four sentences: LENNY, WHERE, WHEN, and YO-YO. Group 1 heard the first two sentences in AMTVS format and the second two sentences in CAMTVS format, and Group 2 heard the first two sentences in CAMTVS format and the second two sentences in AMTVS format. Each of the four sentences was presented three times with a 4-sec ISI. The subjects were then given a 30-sec response interval in which to write down the sentence they heard. Beforehand, the subjects were told that the sentences might be quite difficult and that they should make their best guesses if they were uncertain about a sentence.

In the naturalness rating phase of the experiment, the subjects were presented with the sentences WHERE, LENNY, and WHEN in TVS, AMTVS, and CAMTVS formats. The sentences were presented randomly and repeated twice on each trial. The subjects were required to select a number from 1 to 5 to rate the naturalness of each sentence. The ISI was 2 sec, and the response interval was 10 sec. The instructions were identical to those used in Experiment 1.

Results

The second experiment demonstrated that AMTVS sentences were identified with greater accuracy and were rated more natural than CAMTVS sentences. As in Experiment 1, the subjects' orthographic responses were converted to phonemes before scoring. The mean phonetic accuracy for the AMTVS sentences was 20% greater than that of the CAMTVS sentences. The mean phonetic accuracy scores, broken down by sentence and modulation type, are shown in Figure 6. Comparing across Ex-

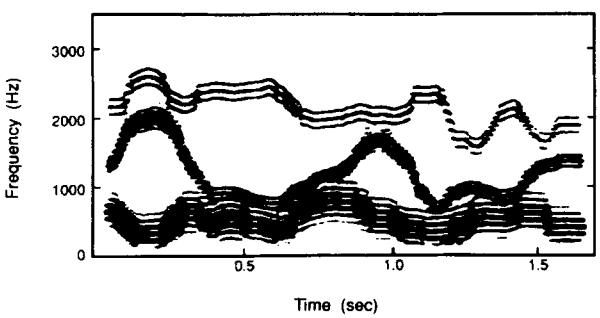

Figure 5. Narrowband spectrogram of the conflicting-rate amplitude-modulated time-varying sinusoidal sentence: "A yellow lion roared."

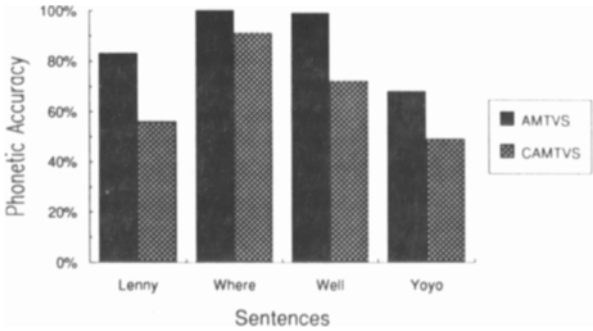

Figure 6. The intelligibility of comodulated versus conflicting-rate-modulated time-varying sinusoidal sentences from Experiment 2.

periments 1 and 2, intelligibility was best with AMTVS sentences and worse with CAMTVS and TVS sentences.

An independent *t* test was performed on this score for each sentence between the AMTVS and CAMTVS conditions. All four of the sentences showed a significant difference between modulation types [$t(31) = 6.42$, $p < .0001$, for LENNY; $t(31) = 2.09$, $p < .05$, for WHERE; $t(31) = 5.82$, $p < .0001$, for WELL; and $t(31) = 3.23$, $p < .003$, for YO-YO]. In every case, the AMTVS sentences were more intelligible than the CAMTVS sentences.

The mean naturalness ratings of each sentence type are shown in Table 2. The AMTVS sentences were rated more natural than the CAMTVS sentences. A planned comparison (Keppel, 1973) between these two sentence types revealed that AMTVS sentences were rated more natural than CAMTVS sentences [$F(1,32) = 66$, $p < .0001$]. A second planned comparison showed that AMTVS sentences were also rated more natural than unmodified TVS sentences [$F(1,32) = 11.59$, $p < .002$]. A final planned comparison showed that the CAMTVS sentences were no more natural than the pure TVS sentences [$F(1,32) = 3.99$, $p > .05$].

Discussion

As predicted, the identification results showed that AMTVS sentences were perceived more accurately than were CAMTVS sentences. That is, the amplitude *comodulation* of the individual components, as opposed to their independent modulation, was critical for improving sentence perception. Unfortunately, interpretation of the intelligibility results is made more difficult by the naturalness findings. Recall that naturalness was predicted to remain high for the CAMTVS sentences because they preserved many of the characteristics of natural speech. Despite this, the listeners rated the AMTVS as more natural sounding than the CAMTVS sentences. Apparently, amplitude comodulation, rather than independent modulation, is also required for naturalness improvements in TVS sentences. In summary, the results from Experiment 2 left both the naturalness-based and auditory-object-based explanations of the comodulation intelligibility improvement as possibilities.

Table 2
Mean Naturalness Ratings for TVS, AMTVS, and CAMTVS Sentences in Experiment 2

Modulation Type	Sentences		
	Where	Lenny	When
No modulation (TVS)	2.73	1.72	2.84
Amplitude modulation (AMTVS)	3.40	2.30	2.93
Conflicting modulation (CAMTVS)	2.83	1.60	2.26

EXPERIMENT 3

Comodulation masking release (CMR) has been argued to be one basis for auditory object formation (Hall, 1987; Yost & Sheft, 1989). It has also been claimed to contribute to pattern analysis in speech perception (Hall & Haggard, 1983). Experiment 3 was designed to directly test whether the improved perception of AMTVS sentences found in Experiments 1 and 2 was due to a comodulation-based grouping effect similar to that found in CMR. If a mechanism similar to the one that underlies CMR were shown to be directly responsible for the improved intelligibility of the AMTVS sentences, then no recourse to a naturalness-based explanation would be necessary.

The hypothesis that the comodulation intelligibility improvement was based on the same auditory principles as CMR was tested by examining the intelligibility of AMTVS sentences at several different modulation rates. Because the CMR effect has been found to be strongest at low modulation frequencies and weakest at higher modulation frequencies (Hall, 1987), intelligibility increments based on CMR should be greatest at low modulation frequencies and least at high modulation frequencies. However, naturalness-based explanations would predict that amplitude comodulation should always improve intelligibility as long as the modulation rate was within the range of human fundamental frequencies.

A set of sentences was created with amplitude comodulation at 50, 100, and 200 Hz, and with no modulation. According to the CMR-based explanation, improved intelligibility due to grouping effects would be expected at 50 and 100 Hz but not at 200 Hz. On the other hand, a naturalness-based explanation would be favored if all three amplitude-modulated conditions were more intelligible than the no-modulation condition. The naturalness-based explanation would also predict that the 100- and 200-Hz modulation conditions would be better perceived than the 50-Hz condition because those modulation rates are more prevalent in voiced speech among the population at large (Peterson & Barney, 1952).

Method

The method was the same as that used in Experiments 1 and 2 with the following exceptions.

Subjects. Thirty-six students were recruited from courses in the Department of Communication Sciences and Disorders at Northwestern University. They had a mean age of 19.8 and ranged from 18 to 29 years of age. The subjects received course credit for their participation in this 20-min experiment.

Stimuli. All stimuli were based on TVS versions of the sentences LENNY, WHERE, YELLOW, and YO-YO. Four sets of test sentences were created, the original TVS sentences plus three AMTVS versions. The AMTVS sentences were amplitude modulated at 50, 100, and 200 Hz with the same methods used in Experiments 1 and 2.

Procedure. The 36 subjects were divided into four groups. Group 1 was presented with LENNY in TVS format, WHERE in 50-Hz AMTVS format, YELLOW in 100-Hz AMTVS format, and YO-YO in 200-Hz AMTVS format. The other three groups contained different pairings of sentence and modulation type to provide complete counterbalancing. Each sentence was presented at each modulation rate over the course of the entire experiment.

As in the previous experiments, the subjects were initially familiarized with each form of modulated sentence they would be hearing during the experiment. To this end, they were presented with three repetitions each of the sentence WELL in TVS format and 50-, 100-, and 200-Hz AMTVS formats. They were informed before the familiarization phase that they would be hearing various versions of the sentence "When were you well?".

Results and Discussion

It was found that the AMTVS sentences modulated at 50 and 100 Hz were generally more intelligible than sentences modulated at 200 Hz or the unmodulated TVS sentences. This is illustrated in Figure 7, which shows the effect of modulation rate on phonetic intelligibility for each of the four sentences. The pattern of results described above was found for three of the four sentences (not WHERE). As in Experiment 1, this sentence was perceived so accurately in all conditions that a ceiling effect apparently masked the results that were found with the other sentences.

A planned comparison was performed on each of the sentences to test the prediction that the 50- and 100-Hz AMTVS sentences would be more accurately perceived than the TVS or 200-Hz AMTVS sentences. The comparison between the 50- and 100-Hz AMTVS sentences and the TVS and 200-Hz AMTVS sentences showed a significant difference between the two sentence types for the sentences LENNY, YELLOW, and YO-YO [$F(1,32) = 44.8$, $p < .0001$; $F(1,32) = 42.8$, $p < .0001$; and $F(1,32) = 73.1$, $p < .0001$, respectively]. It was not significant for WHERE [$F(1,32) = .994$, $p > .3$]. These results are precisely what would be expected if the intelligibility gains of amplitude modulation were related to CMR. That is, those sentences in which the mechanism underlying CMR might be expected to aid in auditory object formation were perceived more accurately than those in which it did not apply. Furthermore, those sentences in which naturalness was predicted to improve intelligibility were not consistently more intelligible.

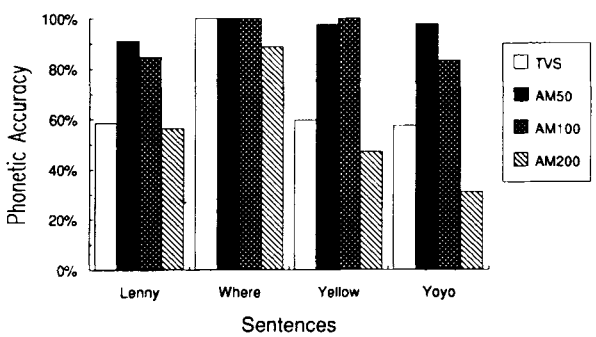

Figure 7. The effect of amplitude modulation rate on the intelligibility of time-varying sinusoidal sentences from Experiment 3.

GENERAL DISCUSSION

The results of all three experiments showed that the amplitude modulation of TVS sentences improved their intelligibility. It was further found that amplitude *comodulation* of each of the tones was required for this improvement; independent amplitude modulation of each of the three component tones was not sufficient. The conditions necessary for improved intelligibility were further refined by showing that comodulation at 50 and 100 Hz increased the intelligibility of TVS sentences, but comodulation at 200 Hz did not.

Taken together, these results were consistent with the hypothesis that a comodulation-based grouping mechanism similar to that underlying CMR was the cause of the improved intelligibility of the AMTVS sentences. The comodulation intelligibility improvement was found here for sentences modulated at frequencies similar to those that produce CMR in psychoacoustic tasks, and it was not found at a frequency where CMR has not been reported. Specifically, Hall and Haggard (1983) reported a maximum CMR effect at comodulation rates of 4 Hz dropping to a minimum (but still present) effect at 64 Hz. Buus (1985) demonstrated CMR at comodulation rates of 15, 50, and 160 Hz, but not at 320 Hz.

Although it proved to be impossible to isolate the effects of amplitude comodulation on naturalness and intelligibility, it was possible to rule out naturalness as the direct cause of the intelligibility-enhancing effect of amplitude comodulation. If the intelligibility improvement were based on increased naturalness, then the 200-Hz comodulation rate should have been as effective as the 100-Hz rate, and perhaps better than the 50-Hz rate because 200 Hz is a very common fundamental frequency for female speakers. Nevertheless, there was no improvement at 200 Hz. Furthermore, pilot experiments in our laboratory have indicated that even when TVS sentences were created with female formant spacings rather than male (to create a more natural pseudofundamental and formant pairing), the sentences were more intelligible at a modulation rate of 100 Hz than 200 Hz. Therefore, while naturalness was correlated with intelligibility, improvements in naturalness were not the cause of improvements in intelligibility.

In natural environments, there can be little doubt that many processes are functioning in concert to extract a desired speech signal from its background. With the present work, amplitude comodulation now joins fundamental frequency and the synchrony of harmonic onset and offset (Darwin, 1981, 1984) as an important cue

to the process of auditory object formation in speech.[3] The perceptual significance of amplitude comodulation at rates of 50–100 Hz has now been established for sentence-length speech signals.

The low-level cues mentioned above are also accompanied by many higher level processes that have been hypothesized to aid fluent speech perception in less-than-ideal listening situations. For example, listener expectations furnish some of the structure by providing a priori constraints on the wide range of possible auditory inputs. Research with time-varying sinusoidal replicas of natural sentences has shown that listeners are able to use high-level expectations to group together acoustically unrelated tones into speech (Best, Studdert-Kennedy, Manuel, & Rubin-Spitz, 1989; Remez et al., 1981). In these experiments, it was found that subjects who knew ahead of time that they would be hearing speechlike sounds performed much better in identification tasks than those who were simply told that they would be hearing computer sounds. Other examples of higher level processes that allow listeners to extract fluent speech from noise include intelligibility improvements based on semantic and syntactic context effects (Grosjean, 1980; Miller, Heise, & Lichten, 1951) and crossmodal effects (McGurk & MacDonald, 1976; Summerfield, 1979).

Because so many grouping processes are available to listeners, their study is very difficult. Further experiments will be required to determine the effect of comodulation on various acoustic foregrounds and backgrounds. In addition, although a strong grouping effect was found in the present experiments, evidence for its use in the segregation of a foreground speech signal from a noise background was not directly addressed. Such an investigation is currently in progress in our laboratory (Carrell, 1990), and similar investigations have been undertaken by others using speech (Grose & Hall, 1992; Scheffers, 1983; Zwicker, 1984) and speechlike (Wright, 1990) stimuli. Finally, a closer examination of the relationship between the intelligibility improvements reported here with psychoacoustic phenomena such as comodulation masking release, modulation detection interference, and comodulation difference detection must be conducted in order to strengthen and refine the proposal that these phenomena are related to the results reported here.

REFERENCES

Best, C. T., Studdert-Kennedy, M., Manuel, S., & Rubin-Spitz, J. (1989). Discovering phonetic coherence in acoustic patterns. *Perception & Psychophysics*, **45**, 237-250.

Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge: MIT Press.

Bregman, A. S., Abramson, J., Doehring, P., & Darwin, C. J. (1985). Spectral integration based on common amplitude modulation. *Perception & Psychophysics*, **37**, 483-493.

Buus, S. (1985). Release from masking caused by envelope fluctuations. *Journal of the Acoustical Society of America*, **78**, 1958-1965.

Carrell, T. D. (1990). The effect of amplitude modulation on the intelligibility of sentences in noise. *Journal of the Acoustical Society of America*, **88**, S174.

Cohen, M. F., & Schubert, E.D. (1987). The effect of cross-spectrum correlation on the detectability of a noise band. *Journal of the Acoustical Society of America*, **81**, 721-723.

Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Quarterly Journal of Experimental Psychology*, **33A**, 185-208.

Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. *Journal of the Acoustical Society of America*, **76**, 1636-1647.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope*, **58**, 955-991.

Gardner, R. B., Gaskill, S. A., & Darwin, C. J. (1989). Perceptual grouping of formants with static and dynamic differences in fundamental frequency. *Journal of the Acoustical Society of America*, **85**, 1329-1337.

Goldstein, J. L. (1973). An optimum processor theory for the central information of the pitch of complex tones. *Journal of the Acoustical Society of America*, **54**, 1496-1516.

Grose, J. H., & Hall, J. W. (1992). Comodulation masking release for speech stimuli. *Journal of the Acoustical Society of America*, **91**, 1042-1052.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, **28**, 267-283.

Hall, J. W. (1987). Experiments on comodulation masking release. In W. A. Yost & C. S. Watson (Eds.), *Auditory processing of complex sounds* (pp. 57-66). Hillsdale, NJ: Erlbaum.

Hall, J. W., & Haggard, M. P. (1983). Co-modulation: A principle for auditory pattern analysis in speech. *Proceedings of the 11th International Congress on Acoustics*, **4**, 69-71.

Hall, J. W., Haggard, M. P., & Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America*, **76**, 50-56.

Keppel, G. (1973). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice-Hall.

Kewley-Port, D. (1976). *A complex-tone generating program*. Research on Speech Perception Progress Rep. No. 3. Bloomington: Indiana University, Department of Psychology, Speech Research Laboratory.

McAdams, S. (1984). The auditory image: A metaphor for musical and psychological research on auditory organization. In W. R. Crozier & A. J. Chapman (Eds.), *Cognitive processes in the perception of art* (pp. 289-323). Amsterdam: North-Holland.

McFadden, D. M. (1987). Comodulation detection differences using noise-band signals. *Journal of the Acoustical Society of America*, **81**, 1519-1527.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **254**, 746-748.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, **41**, 329-335.

Moore, B. C. J. (1989). *An introduction to the psychology of hearing*. London: Academic Press.

Moore , B. C. J. (1990). Co-modulation masking release: Spectro-temporal pattern analysis in hearing. *British Journal of Audiology*, **24**, 131-137.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-950.

Remez, R. E., & Rubin, P. E. (1990). On the perception of speech from time-varying acoustic information: Contributions of amplitude variation. *Perception & Psychophysics*, **48**, 313-325.

Scheffers, M. T. M. (1983). *Sifting vowels: Auditory pitch analysis and sound segregation*. Unpublished doctoral dissertation, Groningen University, The Netherlands.

SCHOONEVELDT, G. P., & MOORE, B. C. J. (1987). Comodulation masking release as a function of signal frequency, flanking band frequency, masker bandwidth, and flanking band level. *Journal of the Acoustical Society of America*, **82**, 1944-1956.

SUMMERFIELD, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, **36**, 314-331.

WRIGHT, B. A. (1990). Comodulation detection differences with multiple signal bands. *Journal of the Acoustical Society of America*, **87**, 292-303.

WRIGHT, B. A., & MCFADDEN, D. (1990). Uncertainty about the correlation among temporal envelopes in two comodulation tasks. *Journal of the Acoustical Society of America*, **88**, 1339-1350.

YOST, W., & SHEFT, S. (1989). Across critical band processing of amplitude modulated tones. *Journal of the Acoustical Society of America*, **85**, 848-857.

YOST, W. A., SHEFT, S., & OPIE, J. M. (1989). Modulation interference in detection and discrimination of amplitude modulation. *Journal of the Acoustical Society of America*, **86**, 2138-2147.

ZWICKER, U. T. (1984). Auditory recognition of diotic and dichotic vowel pairs. *Speech Communication*, **3**, 265-277.

NOTES

1. Speech waveforms contain two primary sources of amplitude modulation, a relatively slow one based on syllable rate (often called the speech envelope) and a faster one based on the fundamental frequency. Both are likely to contribute to auditory object formation, and they rely on different detection mechanisms. Only the more rapid, fundamental-based modulation is considered in the present paper.

2. Precisely how the spectral components are resolved (or processed, if not resolved) is not reviewed here. However, it should be noted that different processes are employed depending on whether or not individual components fall within or across critical bands. For example, at low modulation rates typical of many CMR studies, the modulation is preserved at the output of even very narrow critical bands, whereas at higher modulation rates (similar to male fundamental frequencies), the modulation would not be directly represented in the outputs of many of the critical band filters. Despite this, CMR has been demonstrated at comodulation ranging from several hertz up to 160 Hz (Buus, 1985). In a summary of the CMR literature, Moore (1990, p. 135) concluded that "CMR does not depend on any single cue or mechanism. Rather it reflects the operation of flexible mechanisms which can exploit a variety of cues or combinations of cues depending on the specific stimuli used." Thus, although CMR is used to explain certain findings in the present context, it must be kept in mind that CMR itself is built upon a number of different processes.

3. However, note that Gardner, Gaskill, and Darwin (1989) have found that amplitude comodulation with a dynamically changing rate does not affect phonetic grouping in monosyllables.

(Manuscript received August 1, 1991;
revision accepted for publication March 25, 1992.)