# The Relation Between Speech Intelligibility and the Complex Modulation Spectrum

*Steven Greenberg*[1] *and Takayuki Arai*[2]

[1]1nternational Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA
steveng@icsi.berkeley.edu

[2]Department of Electrical and Electronics Engineering,
Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, Japan
arai@sophia.ac.jp

## Abstract

The amplitude and phase components of the modulation spectrum were dissociated in order to ascertain the importance of cross-spectral, envelope-modulation phase information for understanding spoken language. The dissociation was effected via local time reversals of the speech waveform (i.e., flipping the signal on its horizontal axis) at intervals ranging between 0 and 180 ms. Intelligibility declines progressively as the length of the time-reversed segment increases, down to an asymptotic trough in performance at 100 ms (4% of the words correct). Intelligibility does not correlate highly with the amplitude component of the modulation spectrum, but does coincide closely with the contour of the complex modulation spectrum, a representation that integrates the cross-spectral modulation phase and the conventional (amplitude-based) modulation spectrum into a unified representation. The results imply that intelligibility is based on both the phase and amplitude components of the modulation spectrum.

## 1. Introduction

Speech is highly intelligible under a wide range of acoustic conditions (e.g., reverberation, background noise, competing speakers), even in the presence of significant spectral and temporal distortion. The acoustic basis for this perceptual robustness remains unclear, but is probably related in some measure to the low-frequency modulation spectrum, as suggested by Houtgast and Steeneken many years ago [7]. In their study reverberation sufficiently severe as to impair intelligibility was typically associated with a dramatic alteration of the modulation spectrum. In non-reverberant conditions the modulation spectrum exhibits a peak at ca. 4 Hz (close to the average frequency of a syllable, cf. [1] and [5]), with appreciable energy between 3 and 8 Hz. This “canonical” modulation spectral contour changes significantly under reverberant conditions. The peak lowers to ca. 2 Hz (or less), and the magnitude of this primary locus of energy is appreciably attenuated relative to the unreverberated signal, akin to low-pass filtering of the modulation spectrum.

More recently, Drullman and associates have demonstrated that systematic manipulation of the envelope-modulation characteristics (by either low- or high-pass filtering) results in serious degradation of intelligibility when the integrity of the modulation spectrum in the 3-8 Hz region is compromised [3][4]. However, in that study the modulation spectrum was not differentially filtered across the (tonotopically organized) frequency spectrum, and therefore it was not possible to ascertain if the *phase* component of the modulation spectrum plays an important role in preserving intelligibility.

In order to address this key issue of modulation phase, Greenberg and colleagues created an exceedingly sparse spectral representation of sentential material (with over 80% of the spectrum removed) using1/3-octave channels (spectral “slits”) distributed across the frequency spectrum [6]. When four slits were concurrently presented (slit 1 centered at ca. 335 Hz, slit 2 at ca. 850 Hz, slit 3 at ca. 2135 Hz and slit 4 at ca. 5380 Hz) intelligibility performance was ca. 90% (the “baseline” condition). Shifting the two central slits relative to the lateral channels resulted in a systematic degradation of intelligibility [6]. A shift of 25 ms reduced intelligibility to ca. 80%, and further increases in spectral asynchrony dramatically degraded the ability to understand the sentential material, so that when the central slits were delayed by 75 ms or more, the ability to understand the material was worse than when these same channels were presented in isolation (i.e., there was interference between the central and lateral slits) [10]. Such results suggest that the phase properties of the envelope-modulation pattern across the frequency spectrum may be of importance for understanding spoken language (as spectral asynchrony is equivalent to a change in the modulation spectrum phase across channels).

Because the stimuli in these earlier studies were narrowband in nature, and because the phase of the modulation patterns was not directly measured (but only inferred through a specified degree of asynchrony) it was not possible to conclude with assurance that the phase component of the modulation spectrum is truly the parameter governing the intelligibility of speech. Moreover, the importance of modulation phase may have been overestimated due to a concentration of modulation information in just a few channels of the frequency spectrum. The current study addresses these issues by dissociating the phase and amplitude components of envelope-modulation patterns for full-spectral-bandwidth (6 kHz) signals in order to ascertain whether modulation phase is truly an important parameter for speech intelligibility.

## 2. Stimulus Processing and Presentation

The technique used to dissociate the phase and amplitude components of the modulation spectrum involves local time-reversal of successive segments of the signal waveform (i.e., flipping the waveform on its horizontal axis over a delimited time interval, as illustrated in Figure 1) in a manner similar to that employed (for very different purposes) in [9]. The duration of the reversed segments ranged between 20 and 180 ms (spe-

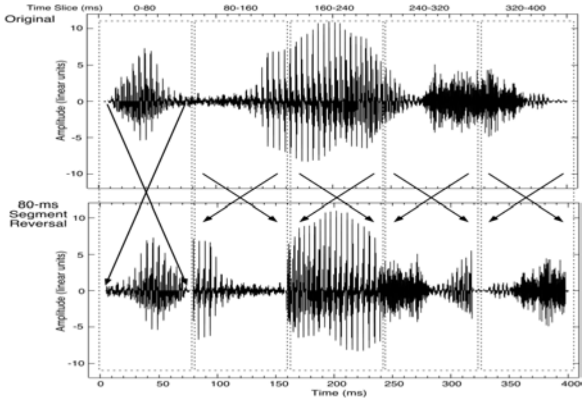


**Figure 1** The stimulus-processing procedure, as illustrated for a brief portion of a single sentence. Each segment was "flipped" on its horizontal axis, preserving all other temporal properties of the signal. In this example the reversed-segment duration is 80 ms. The spectro-temporal consequences of the manipulation are illustrated in Figure 2.

cifically, 20, 40, 50, 60, 70, 80, 100, 140, 180 ms). In addition, an unaltered version of the signals (i.e., a 0-ms condition) was also presented as a control to insure that the original sentences were completely intelligible.

Stimulus materials were derived from the TIMIT corpus (a phonetically balanced set of English sentences with a relatively low degree of semantic predictability) read by speakers (of both genders) spanning a wide range of American dialect regions, chronological ages and voice quality. The acoustic signals were initially sampled at 16 kHz but were low-pass filtered at 6 kHz and quantized with 16-bit resolution.

No sentence was presented on more than a single trial (thus listeners were presented with four different sentences per condition) in order to minimize the effects of learning and memorization on intelligibility performance (in contrast to the study described in [9] where subjects listened to a *single* sentence repeatedly, rating the *subjective* intelligibility of the material). Subjects were allowed to listen to any given sentence up to four times before typing in the word sequence, and were paid for their time. All 27 listeners were native speakers of American English with no reported history of hearing loss. The signals were presented by computer and played over high-quality headphones at a comfortable listening level (under subject control) in a sound-attenuated room.

## 3. Data Collection and Analysis

Each subject listened to five practice sentences before beginning the experiment proper. Listeners were instructed to type the sequence of words heard directly into the computer. Although each subject listened to the same sentences as presented to the other listeners, the specific sequence of sentences played (in terms of the specific association of sentence material with time-reversed-segment duration) was varied so as to minimize the likelihood that variation in intelligibility could be attributed to the identity of specific sentential material. The number of correct words per sentence was scored using an algorithm that automatically compensated for minor errors in spelling. The proportion of words correctly typed (and in the proper sequence) was computed for each set of sentences associated with a specific experimental condition.

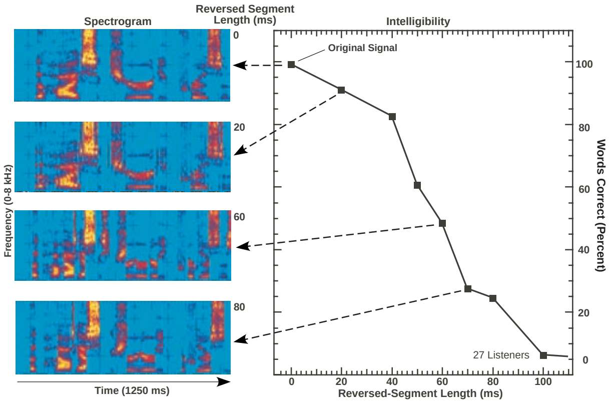


**Figure 2** Intelligibility of TIMIT sentences as a function of reversed-segment duration (right panel). Each data point represents an average of 27 listeners. Spectrograms associated with a single sentence from the corpus are illustrated for three different reversed-segment durations (20, 60 and 80 ms), as well as for the original signal (0 ms).

## 4. Intelligibility of Locally Time-Reversed Speech

The average intelligibility of locally time-reversed sentences is shown in Figure 2. Reversed-segment intervals of 40 ms or less result in relatively modest levels of intelligibility degradation relative to the original signal. For longer reversed-segment durations the decline in intelligibility is precipitous and unrelenting. For reversal intervals of 100 ms or longer the intelligibility is close to zero (6% at 100 ms, 4% at 140 ms and 3.5% at 180 ms - the latter two conditions are not plotted for clarity of illustration).

Spectrograms associated with reversed-segment durations of 20, 60 and 80 ms for a single sentence are shown in Figure 2 (along with a spectrogram of the original signal). The spectra of the 60- and 80-ms time-reversed signals are highly distorted relative to the original, and therefore it is not surprising that their intelligibility is markedly degraded (48% and 24%, respectively). It is of interest to ascertain whether a specific property of the modulation spectrum (particularly in the key 3–

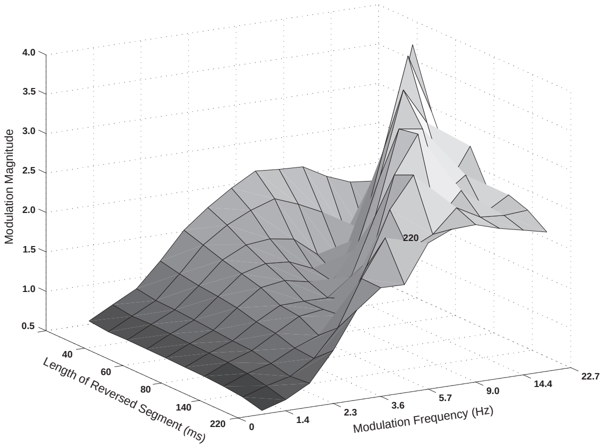


**Figure 3** The *amplitude component* of the modulation spectrum computed for all 40 sentences used in the experiment as a function of reversed-segment duration. Note that there is a slight decline in magnitude in the key, 3–9 Hz region for reversed-segment lengths of 20–50 ms, followed by a steep *increase* in magnitude for longer reversed-segment intervals.

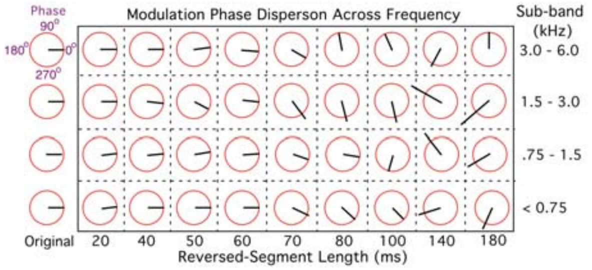


**Figure 4** Modulation phase (the *magnitude* of the phase vector is indicated by the length of the line in the phasor circle) dispersion (for a single frequency, 4.5 Hz) across the frequency spectrum as a function of reversed-segment duration for a single sentence. The frequency spectrum is partitioned into three one-octave sub-bands plus a fourth sub-band for energy below 750 Hz. Note that the phase dispersion across sub-bands increases markedly for reversed-segment durations greater than 60 ms.

8 Hz region) is correlated with the systematic decline in intelligibility as the length of the time-reversed segment increases.

## 5. The Modulation Spectrum (Amplitude Component)

The amplitude component of the modulation spectrum was computed for all 40 sentences serving as stimuli in the experiment (Figure 3). The amplitude component was computed at 13 center frequencies (1.1, 1.4, 1.8, 2.3, 2.9, 3.6, 4.5, 5.7, 7.2, 9.0, 11.4, 14.4, 18.1 Hz). The bandwidth of each modulation spectral channel was 1/3 of an octave (at the half-power points, similar to the filter bandwidths used by Houtgast and Steeneken [7]). Each signal was decimated by a factor of 200 (i.e., the effective sampling rate was 80 samples per second) and subsequently low-passed filtered (using a linear-phase FIR filter) at 30 Hz. The modulation spectrum shown in Figure 3 was computed for the entire (6-kHz) bandwidth of the stimulus material.

The amplitude component of the modulation spectrum does not correlate well with the intelligibility data of Figure 2. Initially, there is a steep decline in amplitude in the 3–9 Hz region of the spectrum for time-reversed segments up to 50 ms, followed by a steep *increase* in magnitude for segment reversals of longer duration. In contrast, the gradient of intelligibility declines progressively with increasing length of the time-reversed segment interval. Thus, the conventional form of the modulation spectrum does not appear to predict with precision the intelligibility associated with locally time-reversed signals.

## 6. Modulation Phase Dispersion Across Frequency

Two earlier studies (described in Section 1) implied that intelligibility may depend on the phase of the modulation patterns distributed across the frequency spectrum [6][10]. This assumption was tested in preliminary fashion by examining the phase pattern associated with the envelope-modulation patterns across (tonotopic) frequency for a single sentence (and at a single tap, 4.5 Hz, close to the peak of the modulation spectrum), as illustrated in Figure 4. The spectrum was partitioned into four sub-bands, the three highest of which were an octave wide. The lowest sub-band encompassed frequencies below 750 Hz.

The phase of the 4.5-Hz component of the modulation spectrum is, by definition, coherent (and set to $0^o$) for the original signal (left-most column in Figure 4). For time-reversal intervals of short duration (20-40 ms) the phase dispersion (relative to the original signal) is small. There are no apparent phase shifts of more than $5^o$ in these conditions. As the length of the time-reversed segment increases to 50 and 60 ms, the phase dispersion increases to ca. 15%. For longer time-reversal intervals the phase dispersion across the frequency spectrum increases even further. The pattern of cross-spectral modulation phase dispersion observed for this *single* sentence and modulation frequency roughly parallels the intelligibility gradient in Figure 2. Is this phase pattern representative of the material as a whole? And if so, does it occur across the key 3–8 Hz range of the modulation spectrum?

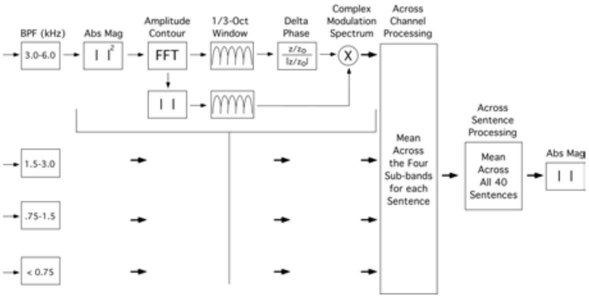

BPF (kHz)
3.0-6.0
1.5-3.0
.75-1.5
< 0.75
Abs Mag
| |²
Amplitude Contour
FFT
| |
1/3-Oct Window
Delta Phase
z/z₀ / |z/z₀|
Complex Modulation Spectrum
X
Across Channel Processing
Mean Across the Four Sub-bands for each Sentence
Across Sentence Processing
Mean Across All 40 Sentences
Abs Mag
| |

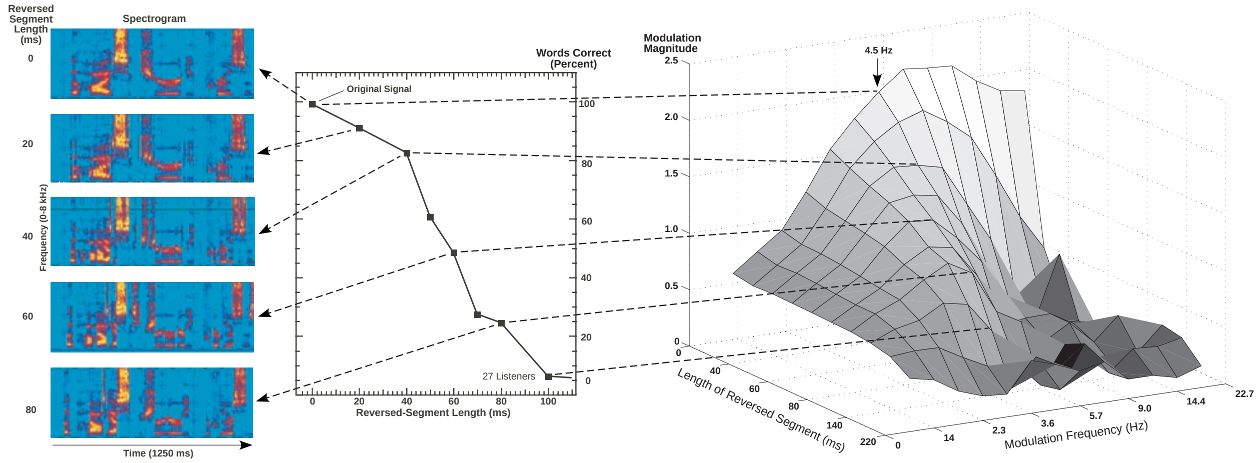
**Figure 5** Signal-processing procedure for computing the complex modulation spectrum of the sentential material. The magnitude and phase components of the modulation spectrum are initially computed separately for each of the four sub-bands. The delta-phase (relative to the original signal) is computed for each one-third-octave interval of the modulation spectrum and then combined with the commensurate amplitude component to obtain the *complex* modulation spectrum (phase and amplitude combined) as illustrated in Figure 6.

## 7. The Complex Modulation Spectrum

In order to ascertain whether the modulation-phase-dispersion pattern observed in Figure 4 is representative of the experimental stimuli as a whole, a novel method was developed for combining the phase and amplitude components of the modulation spectrum into a single representation.

This “complex” modulation spectrum initially computes the amplitude and phase components for each of the 13 modulation spectral taps in segregated fashion, as shown in Figure 5. The amplitude component is computed in a manner similar to that described in Section 5, the key difference being that the computation is performed for each of the four sub-bands separately, rather than over the entire spectrum.

The computation of the phase component is referenced to the phase of the original signal for each tap-point of the modulation spectrum (and for each sub-band of the frequency spectrum). This step essentially corresponds to a computation of the delta phase (relative to the original, perfectly intelligible, signal) and is merely a means with which to quantitatively track changes in the phase pattern associated with key regions of the modulation spectrum across the four spectral sub-bands. If the phase is coherent across the frequency spectrum, then the magnitudes associated with the amplitude component of the modulation spectrum will sum in linear fashion. This is clearly what occurs when the length of the time-reversed segment is less than 50 ms, as illustrated in Figure 6. As the reversed-seg-



**igure 6** Intelligibility of locally time-reversed sentences (center) to the complex modulation spectrum (right) for reversed-segment durations between 20 and 100 ms (and the original signal). Spectrograms of a sample sentence are shown on the left. As the reversed-segment duration increases beyond 40 ms, intelligibility declines precipitously, as does the magnitude of the complex modulation spectrum. The spectro-temporal properties of the signal also degrade significantly under such conditions.

ment interval lengthens, the phase dispersion increases significantly, thereby reducing the component magnitude associated with the combined phase and amplitude information. There is a progressive decline in the magnitude of the complex modulation spectrum as a function of time-reversed-segment duration across all frequencies within the 3–9 Hz region. And the slope of the decline in the complex modulation spectrum in this portion of the spectrum parallels that of the intelligibility function to a remarkable degree.

## 8. Discussion and Conclusions

Although the intelligibility of spoken language depends in some fashion on the integrity of the modulation spectrum in the region between 3 and 8 Hz (e.g., [3][4][6][7]), the precise relation between the envelope-modulation patterns and specific linguistic attributes has yet to be clearly delineated. The results of the current study suggest that the modulation spectrum is important not just for providing a basis for syllabic segmentation (as suggested in [1][2][5]), but is also important for defining finer-grained phonetic information (such as place and manner of articulation). This fine phonetic detail vitally depends on the pattern of cross-spectral modulation, and is the most likely part of the speech signal to be distorted upon alteration of the phase component of the modulation spectrum. For this reason a complete account of the modulation spectrum's significance for speech understanding (as well as for automatic speech recognition by computer, cf. [8]) is likely to require specification of both the amplitude and phase components of this representation.

## 9. Acknowledgements

The research described in this paper was supported by the National Science Foundation and the U.S. Department of Defense. We thank Rosaria Silipo, Joy Hollenback and Shino Sakaguchi for their assistance in running the experiments, as well as the listeners for their diligence and time. The experimental protocol conformed to the guidelines of the Committee for the Protection of Human Subjects at the University of California, Berkeley.

## 10. References

[1] Arai, T. and Greenberg, S. “The temporal properties of spoken Japanese are similar to those of English,” *Proc. Eurospeech*, pp. 1011-1014, 1997.

[2] Arai, T. and Greenberg, S. “Speech intelligibility in the presence of cross-channel spectral asynchrony,” *Proc. IEEE ICASSP*, pp. 933-936, 1998.

[3] Drullman, R., Festen, J.M. and Plomp, R. “Effect of temporal envelope smearing on speech reception.” *J. Acoust. Soc. Am.*, 95: 1053–1064, 1994.

[4] Drullman, R., Festen, J.M. and Plomp, R. “Effect of reducing slow temporal modulations on speech reception, *J. Acoust. Soc. Am.*, 95: 2670–2680, 1994.

[5] Greenberg, S. “Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation,” *Speech Communication*, 29: 159-176, 1999.

[6] Greenberg, S., Arai, T. and Silipo, R. “Speech intelligibility derived from exceedingly sparse spectral information.” *Int. Conf. Spoken Lang. Proc.*, pp. 2803-2806, 1998.

[7] Houtgast, T. and Steeneken, H. “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria.” *J. Acoust. Soc. Am*, 77: 1069-1077, 1985.

[8] Kanedera, N., Arai, T., Hermansky, H. and Pavel, M. “On the importance of various modulation frequencies for speech recognition,” *Proc. Eurospeech*, pp. 1079-1082, 1997.

[9] Saberi, K. and Perrott, D.R. “Cognitive restoration of reversed speech,” *Nature*, 398: 760, 1999.

[10] Silipo, R., Greenberg, S. and Arai, T. “Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations,” *Proc. Eurospeech*, pp. 2687-2690, 1999.