INVITED PAPER *Special Section on Speech Dynamics by Ear, Eye, Mouth and Machine*

What are the Essential Cues for Understanding Spoken Language?

Steven GREENBERG[†], ***Nonmember and*** **Takayuki ARAI**[††a)], ***Member***

SUMMARY Classical models of speech recognition assume that a detailed, short-term analysis of the acoustic signal is essential for accurately decoding the speech signal and that this decoding process is rooted in the phonetic segment. This paper presents an alternative view, one in which the time scales required to accurately describe and model spoken language are both shorter and longer than the phonetic segment, and are inherently wedded to the syllable. The syllable reflects a singular property of the acoustic signal — the modulation spectrum — which provides a principled, quantitative framework to describe the process by which the listener proceeds from sound to meaning. The ability to understand spoken language (i.e., intelligibility) vitally depends on the integrity of the modulation spectrum within the core range of the syllable (3–10 Hz) and reflects the variation in syllable emphasis associated with the concept of prosodic prominence ("accent"). A model of spoken language is described in which the prosodic properties of the speech signal are embedded in the temporal dynamics associated with the syllable, a unit serving as the organizational interface among the various tiers of linguistic representation.

key words: *speech perception, intelligibility, syllables, modulation spectrum, auditory system, auditory-visual integration*

1. Introduction

Models of spoken language have traditionally focused on two disparate aspects of speech recognition — the acoustic-phonetic properties of the speech signal and the psychological processes associated with lexical access [19]. Such traditional models assume that speech can be characterized merely as a sequence of words containing strings of phonemic constituents. Within this traditional "beads-on-a-string" framework the process of lexical recognition is primarily one of decoding the phonemes associated with the speech signal, attempting to deduce the sequence of words uttered by the speaker via a process of phonetic characterization. Such phonetic-segment models possess a beguiling simplicity and directness — word recognition is merely a matter of decoding the sequence of phones uttered and then proceeding to "look up" the word in a mental lexicon, akin to the process by which a reader retrieves an entry in a dictionary [11].

However attractive such sequential models may be in the abstract, they fail to account for many properties of spoken language such as (1) the ability of listeners to understand speech under a broad range of conditions that distort many of the acoustic-phonetic properties of the signal via such interference as reverberation and background noise, and (2) the remarkable degree of pronunciation variation observed at the phonetic level in everyday speech [10]. These two signatures of "real-world" speech — acoustic distortion and pronunciation variability — render untenable many contemporary models of speech recognition, as well as wreak havoc with current-generation automatic (machine) speech recognition systems. An alternative theoretical formulation is required, one capable of accounting for the patterns of pronunciation variation in casual speech [10], [11], and that provides a principled mechanism for the stability of intelligibility under the broad constellation of acoustic conditions characteristic of the real world. Moreover, the theoretical formulation should be capable of accounting for the importance of visible speech information, particularly under conditions of acoustic interference and non-native familiarity with the language spoken.

Manuscript received October 9, 2003.

[†]The author is with The Speech Institute, Santa Venetia, CA 94903, USA.

[††]The author is with Sophia University, Tokyo, 102–8554 Japan.

a) E-mail: arai@sophia.ac.jp

2. The Doors of Perception

One means by which to ascertain the essential cues for understanding spoken language is to artificially distort the speech signal and measure its impact on intelligibility. Such distortions may be used to expose chinks in the perceptual armor that normally shield the brain from the deleterious effects of reverberation and other forms of acoustic interference.

The perceptual studies described in this paper are all linked in some fashion to the low-frequency (3–20 Hz) modulation spectrum. The modulation spectrum reflects fluctuations in energy associated with articulatory dynamics pertaining to the movement of the lips, jaw and tongue during the production of speech. The modulation of energy at such low frequencies is inherently tied to the syllable, the linguistic unit most closely associated with articulatory gestures. The duration of syllables varies greatly, and is reflected in the modulation pattern of the speech signal. Moreover, syllable structure, both in terms of phonetic constituents, as well as prosodic prominence, is reflected in syllabic duration and hence in the modulation spectrum.

3. Syllable Duration and the Modulation Spectrum

Syllable duration varies roughly ten-fold in stress-timed languages such as English, ranging between 50 and 500 ms [10]. Even in Japanese, a language noted for its relatively even tempo, syllables vary in length between 50 and

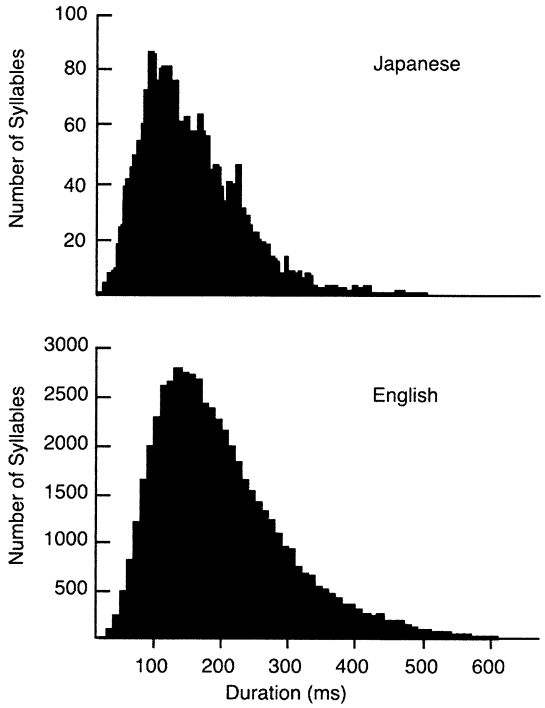


Fig. 1 Statistical distribution of syllable duration for spontaneous material in Japanese and American English. Adapted from [1].

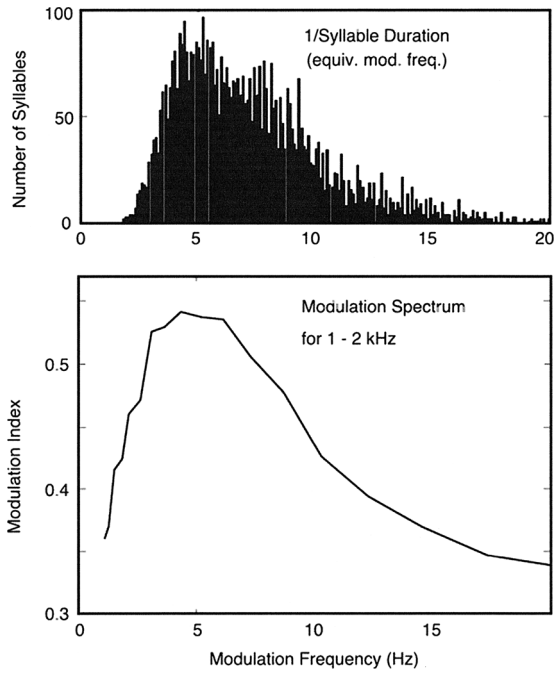


Fig. 2 The relation between the distribution of syllable duration (transformed into modulation frequency) and the modulation spectrum of the same Japanese material as shown in Fig. 1, computed for the octave region between 1 and 2 kHz. Adapted from [1].

300 ms [1]. Figure 1 illustrates the distribution of syllable duration for English and Japanese spontaneous material. Overall, the distributions are similar — the principal characteristic distinguishing the two languages pertains to the proportion of syllables longer than 300 ms. In English, approximately 15% of the syllables are longer than this interval [10], while Japanese contains far fewer syllables (approximately 1–2%) of this length [1]. In this sense the distinction between a stress-timed and syllable (or mora)-timed language is relatively subtle — a matter of the proportion of syllables exceeding a specific interval of time. This difference in duration reflects two specific properties of English syllable structure that is encountered in Japanese to a far lesser degree: (1) a profusion of consonant clusters, particularly at syllable onset, and (2) syllable lengthening, particularly in the vocalic nucleus, associated with prosodic prominence [14], [16].

Figure 2 illustrates the close relationship between syllable duration and the modulation spectrum. The bandwidth of the modulation spectrum is quite broad, encompassing frequencies between 3 and 20 Hz, consistent with the broad variation in syllable duration observed. Because of the intimate relationship between energy dynamics, speech production and syllable structure, there is an intrinsic correlation between syllable duration and the modulation of the signal's waveform. The core of the syllable is the nucleus, which is almost always vocalic and usually voiced. The nucleus contains the greatest overall energy within the syllable, serving as the structural foundation upon which the onset and coda constituents lie. A syllable is only a syllable by virtue of its nucleic core. The onset and coda are optional constituents. In many languages, such as Japanese, the nucleus is the dominant constituent of the syllable. The amplitude of the onset and coda is generally lower than that of the nucleus and conforms to an "energy arc" in which the sound pressure rises and falls gradually over time [11]. This energy gradient is the principal reason why the modulation spectrum of speech is dominated by frequencies in the 3–10 Hz region rather than by components higher than 20 Hz [11].

The relation between syllable structure and the modulation spectrum is manifest in prosodic prominence. In all languages there are certain syllables that are more linguistically and perceptually prominent than others. In stress-timed languages syllable duration (particularly pertaining to the nucleus) plays an important role [4]. However, even in non-stress languages, duration is likely to play some role in the specification of prosodic accent that is reflected in the modulation spectrum.

Figure 3 illustrates the relationship between prosodic prominence, word duration and the modulation spectrum for American English. The material is derived from a corpus of spontaneous telephone dialogues (SWITCHBOARD — [7]) that was phonetically and prosodically annotated by trained linguistic transcribers [10]. In this corpus most words contain only a single syllable, so that word duration and syllable length are largely coterminous. Words without accent (i.e., unstressed) are generally shorter than 200 ms, while heavily accented words are usually longer than this interval (Fig. 3). More importantly, the distributions associated with unaccented and accented lexical forms overlap only to

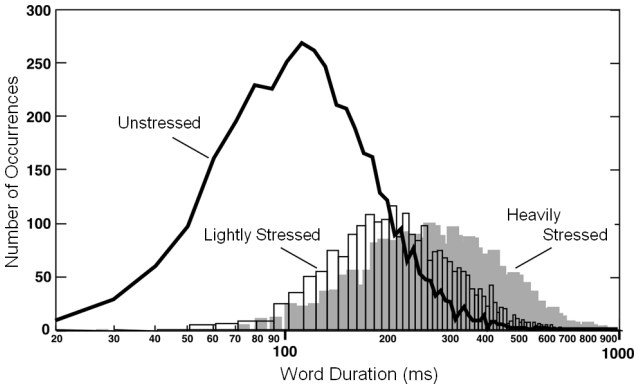


Fig. 3 Word duration as a function of stress-accent level for a corpus of spontaneous American English. Frequency histograms of words (n = 10,001) associated with a range of stress-accent levels are shown. Eighty percent of the words are monosyllablic. For those containing more than a single syllable, a word is deemed as stressed if it contains at least one syllable of that accent level (the most heavily accented syllable in the word determining its stress-accent pattern). Unstressed words are entirely without stress in any syllable. The solid black curve represents the histogram for unstressed words (n = 3946). The histogram associated with lightly stressed words (n = 2484) is represented by unfilled black columns. Heavily stressed words (n = 3571) are shown in grey. The bin width for lexical duration is 10 ms.

a slight degree (Fig. 3), suggesting that these represent two distinct classes (at least with respect to duration).

Given the relation between syllable duration and the modulation spectrum, one may infer that the lower, left-hand branch of the spectrum (< 4 Hz) is largely associated with heavily accented syllables, while the upper branch (> 5 Hz) is most closely associated with unaccented forms. The central core of the modulation spectrum, between 4 and 5 Hz, pertains to syllables spanning a range of accent levels (i.e., represents the convergence of accented and unaccented forms).

4. The Perceptual Significance of the Modulation Spectrum

A variety of studies have shown that intelligibility depends on the integrity of the low-frequency modulation spectrum [2], [3], [5], [13], [21]. Within reverberant environments the greatest impact of acoustic reflections is between 2 and 6 Hz [18]. In an overly reverberant environment the acoustic reflections combine with the original signal to reduce the amount of modulation apparent in the waveform. The consequence of such reflections is to reduce the magnitude of the modulation spectrum at its peak, as well as to reduce the spectral peak from 4–5 Hz to ca. 1–2 Hz. If the amount of modulation attenuation is not too severe the speech signal is still intelligible, if somewhat distorted in quality. However, significant attenuation of the modulation spectrum results in a drastic decline in intelligibility, akin to speech broadcast over a poor-quality public address system as often occurs in a train station or other large enclosure with highly reflective surfaces. Houtgast and Steeneken demonstrated many years ago that the modulation spectrum was a good predictor of speech intelligibility over a broad range of acoustic environments [18]. More recently, Drullman and colleagues have shown that the key portion of the modulation spectrum for intelligibility lies below 8 Hz [5], a result confirmed for Japanese [3] and for English [2]. Thus, it is clear that the integrity of the modulation spectrum is both essential for understanding spoken language and that its general characteristics reflect something important about syllables with respect to their structure and/or segmentation.

But what precisely in the modulation spectrum is so crucial for understanding spoken language? And how can such knowledge be used for developing technology capable of benefiting large numbers of people? Perceptual studies, such as those described below, may provide a certain degree of insight of potential utility for speech technology and scientific knowledge.

5. Spectral Asynchrony's Impact on Intelligibility, and Its Relation to Reverberation

One effect of reverberation is to jumble the spectral content of the acoustic signal across both time and frequency, particularly that portion of the spectrum below 1500 Hz. Although reverberation is known to interfere with intelligibility, particularly among the hearing impaired, the basis for its deleterious impact is not well understood.

In order to gain some insight into reverberation's impact on intelligibility the following experiment was performed. The spectrum of spoken sentences (American English, the TIMIT corpus) was partitioned into 19 quarter-octave channels and quasi-randomly jittered in time. The jittering algorithm insured that adjacent channels were desynchronized by a minimum time interval (at least one-quarter of the maximum interval) and that the mean temporal jitter across channels was equal to precisely one half of the maximum jitter interval. Using such an algorithm provides a convenient means of comparing the amount of cross-spectral jitter imposed with durational properties of important linguistic units such as the phonetic segment and the syllable.

In order to assess the impact of time jitter on the ability to understand spoken language, the amount of (maximum) spectral asynchrony was varied systematically between 20 and 240 ms, and the effect on intelligibility measured (in terms of the proportion of words correctly reported). But this relation alone does not provide much insight into the underlying mechanisms responsible for understanding speech presented under such distorted conditions. For this reason modulation spectra were also computed for the sentence material used. The modulation spectra were computed for delimited spectral regions ("sub-bands") rather than across the entire frequency range (6 kHz) of the acoustic signal. Three of the four sub-bands were an octave wide, while the lowest contained all frequencies below 750 Hz. The highest sub-band encompassed 3.0–6.0 kHz, while the other octave bands ranged between 1.5–3.0 kHz and 0.75–1.5 kHz. In this fashion it was possible to compute the modulation spectrum for each sub-band and relate this pattern to the decline

of intelligibility. In order to provide a simple, quantitative metric relating the modulation spectrum and acoustic frequency to intelligibility, the magnitude of the modulation spectrum in the crucial 3–6 Hz region was computed for each sentence and time-jitter condition. This information was normalized relative to the original, undistorted signal for each sentence and the results plotted on the same scale as the data pertaining to intelligibility for the 27 listeners participating in the experiment.

The results of this study are shown in Fig. 4 (and discussed in greater detail in [2]). Intelligibility declines progressively with increasing amounts of cross-spectral jitter. The surprising aspect of these data concerns the relatively modest impact on intelligibility exerted by significant amounts of cross-spectral time jitter. Even when the spectral asynchrony encompasses a (maximum) range of 140 ms (i.e., a mean jitter interval of 70 ms), 75% of the words are accurately reported by listeners. This jitter interval is equivalent to the average duration of a phonetic segment, and implies that the auditory system (and the brain) is exceedingly tolerant of spectro-temporal distortion in the speech signal.

This tolerance of spectral asynchrony is precisely what one would expect of a processing mechanism that has evolved to decode acoustic signals propagating through environments of variable (and unpredictable) nature. In particular, such robustness to cross-spectral time jitter would be quite useful in reverberant environments, where acoustic reflections are commonplace, and would help to instill some measure of perceptual stability amidst a highly variable background.

The intelligibility data also suggest that listeners rely on different parts of the acoustic spectrum for decoding speech, depending on the specific background conditions. When there is relatively little cross-spectral jitter the data imply that frequencies below 1.5 kHz play a particularly important role in decoding the signal. As the amount of spectral asynchrony increases, the decline in intelligibility closely mirrors the fall-off of the modulation spectrum in the channels above 1.5 kHz. To the extent that such jitter simulates the distortion imposed by reverberation, this result implies that the high-frequency portion of the spectrum plays a particularly important role in decoding speech under reverberant conditions.

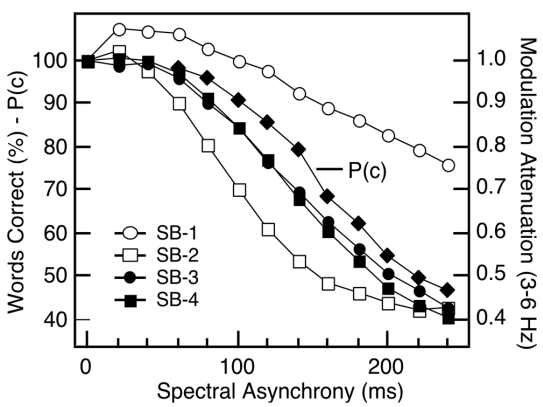


Fig. 4 Intelligibility ($P(c)$) of de-synchronized spectral channels as a function of the amount of maximum cross-spectral asynchrony imposed. This function is compared with the magnitude of the modulation spectrum in the range of 3–6 Hz for four separate sub-bands for the 40 TIMIT sentences used in Experiment 1. From [2].

This spectrally adaptive strategy could be important for understanding the frequency-selective nature of intelligibility deficits among the hearing impaired. In quiet listening conditions such individuals rarely experience a significant problem understanding speech. However, in reverberant and noisy conditions their ability to comprehend declines markedly. The data in Fig. 4 provide a potential explanation for this selective deficit. The effects of reverberation and other common forms of background interference are particularly pronounced below 1.5 kHz. Normal-hearing individuals may shift their listening strategy under such conditions to rely on spectral information above 1.5 kHz that is largely redundant with the low frequencies in quiet. The hearing impaired rarely possess the luxury of such redundant processing capacity given their deficit's concentration in channels above 1.5 kHz. Thus, it is likely that in reverberant conditions, these individuals would be less able to extract useful information from the high-frequency channels to aid in decoding the speech signal. Under such conditions, they may rely on visual cues (i.e., "speechreading") to a large extent (see [8]). In the absence of such visual cues, the hearing impaired are likely to experience extreme difficulty understanding speech associated with reverberant and other noisy environments. Visual supplementation of the acoustics could be extremely important under such circumstances (see Sect. 9).

6. How Much of the Acoustic Spectrum is Required to Understand Spoken Language?

The intelligibility data described in Sect. 5 imply that the auditory system (and by extension, the brain as a whole) is remarkably insensitive to time jitter across the spectrum. However, this assumption was not directly tested in Experiment 1 because of the nature of the signals used. In that study the signal spectrum was 6 kHz wide (and continuous), comprising much of frequency information utilized for speech decoding under favorable listening conditions. In such circumstances there is a considerable redundancy in the signal that listeners may exploit to decode the speech signal. Exploitation of such redundancy could potentially foil the experimental design's intent through temporal correlation of modulation patterns distributed across the (tonotopic) spectrum. If listeners were capable of selectively focusing on just a few channels broadly distributed across the frequency spectrum whose modulation patterns are largely in synch (relative to the original, undistorted signal), the auditory system's "true" sensitivity to cross-spectral jitter may have been grossly underestimated.

A statistical analysis of the cross-spectral time-jitter patterns used in Experiment 1 is consistent with this intuition. If only four channels are chosen from the nine-

teen used in the experiment, and these four are distributed across the frequency spectrum so that one channel comes from the lowest sub-band (< 750 Hz), one from the highest sub-band (3–6 kHz) and the remaining two from the other sub-bands (0.75–1.5 kHz and 1.5–3 kHz), then it is possible to ascertain the time jitter of four such channels relative to each other as a means of estimating the potential magnitude of temporal correlation across frequency. Such an analysis shows that ca. 10% of the 448 potential 4-channel combinations exhibit very little time jitter across the spectrum [13]. Thus, if listeners were able to decode the speech signal from just four channels widely distributed across the acoustic spectrum, and if it were possible to determine which of the 19 channels were synchronized to each other, then it would be possible, in principle, to decode the speech signal through some form of frequency-selective listening.

In order to test this possibility a second perceptual experiment was performed. In contrast to the original set of signals, in which all frequencies below 6 kHz were presented to the listener, sentences were filtered into narrow (1/3-octave) channels and most of the spectral information discarded. Listeners were asked to report the words heard when one, two, three or four channels were presented concurrently (and in synch with each other). The lowest channel was centered at 335 Hz, the second at 850 Hz, the third at 2135 Hz and the fourth at 5400 Hz. Intelligibility depends on the number of channels presented, as well as their position within the frequency spectrum (Fig. 5). The most pertinent result is the intelligibility associated with four concurrently presented channels. Under such conditions approximately 90% of the words are accurately reported. This result is important, as it demonstrates that a detailed spectro-temporal representation of the speech signal is not required for the acoustic signal to be intelligible. Three-quarters of the spectrum was discarded without a significant decline in the ability to correctly identify the words spoken. Note, that for the example shown in Fig. 6, the modulation pattern associated with the four-channel signal is remarkably similar to that of the original waveform (except for a scaling factor), suggesting that sparsely sampling the spectrum in this fashion is capable of capturing the essence of the sentence's modulation properties. To the extent that intelligibility is derived from such modulation patterns, it is perhaps not so surprising that virtually all of the words spoken are accurately reported. The experimental results also provide a stable baseline with which to measure the impact of various signal manipulations (as described below).

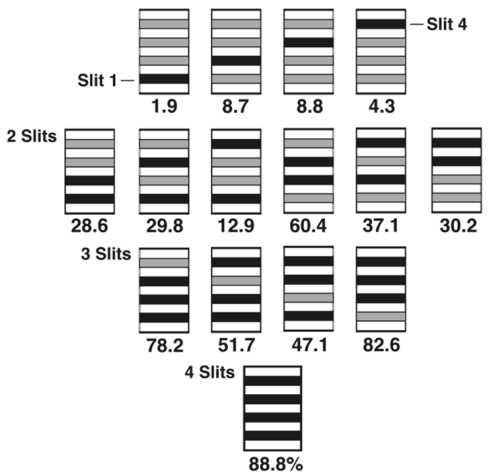


Fig. 5 Intelligibility of sparse-spectrum sentences under 15 separate listening conditions. Baseline word accuracy is 88.8% (4-channel condition). The intelligibility of the multiple-channel signals is far greater than would be predicted on the basis of word accuracy (or error) for individual channels presented alone. The region between 750 and 2400 Hz (slits 2 and 3) provides the most important intelligibility information. From [13].

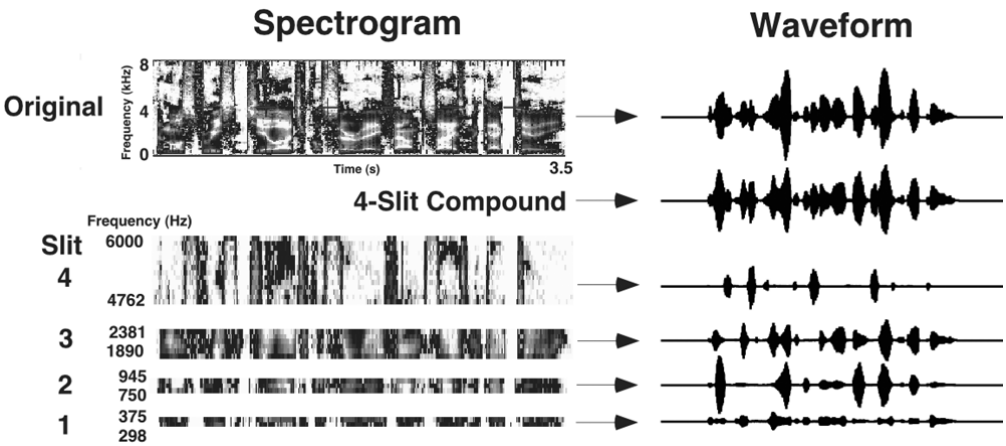


Fig. 6 Spectrographic and time-domain representations of a representative sentence ("The most recent geological survey found seismic activity") used in the study. The channel waveforms are plotted on the same amplitude scale, while the scale of the original, unfiltered signal is compressed by a factor of five for illustrative clarity. The frequency axis of the spectrographic display of the channels has been non-linearly compressed for illustrative tractability. Note the quasi-orthogonal temporal registration of the waveform modulation pattern across frequency channels. From [13].

7. Measuring the Auditory System's "True" Sensitivity to Spectral Asynchrony

Using a performance baseline for various combinations of sparse spectral sentences, one can measure intelligibility when the channels are desynchronized relative to each other — this is the basis of Experiment 3. Listeners were presented only a small proportion of the spectrum with which to decode the sentential material. Because the listening task is intrinsically difficult — listeners are rarely able to accurately report all of the words presented — it encourages subjects to use all of the spectral information provided. Thus, when some of the spectral information is removed or desynchronized relative to other channels, a reliable estimate can be made of the contribution to intelligibility made by each of the channels involved. It is thus possible to estimate the contribution of each portion of the spectrum to overall intelligibility, as well as to measure how sensitive the auditory system is to cross-spectral time jitter imposed on the modulation patterns associated with each spectral channel. For the present discussion we focus on one form of channel asynchrony, in which the two central channels either lead or lag the lateral channels in time (in a separate experiment, also shown, a single central channel was desynchronized relative to its counterparts).

Figure 7 shows how desynchronizing one or both of the two central channels relative to their lateral counterparts affects intelligibility for various degrees of asynchrony. When the channels are jittered by 25 ms the impact on intelligibility is small, ranging between 10 and 20%. For greater amounts of asynchrony the impact is profound — intelligibility falls to 60% or less. Increasing the cross-spectral asynchrony lowers intelligibility for jitter intervals up to ca. 250 ms (Fig. 8). Beyond this limit additional delays result in a slight improvement in intelligibility overall (but is quite variable across listeners and sentential material). Such results imply that the auditory system is quite sensitive to spectral asynchrony — it is necessary only to use a sufficiently sensitive assay in order to expose the chinks in the brain's perceptual armor. Moreover, when asynchrony across channels exceeds 50 ms, intelligibility descends below performance baseline for the two central channels presented alone (ca. 60%), suggesting that there is active interference and that the brain is far less tolerant of cross-spectral asynchrony than implied in Experiment 1.

The apparent tolerance of cross-spectral jitter in the first experiment is probably the result of redundancy among channels that affords numerous opportunities for modulation information to be combined across the spectrum for optimum decoding of linguistic information contained in the speech signal. Such redundancy masks the exquisite sensitivity of the auditory system to cross-spectral time jitter.

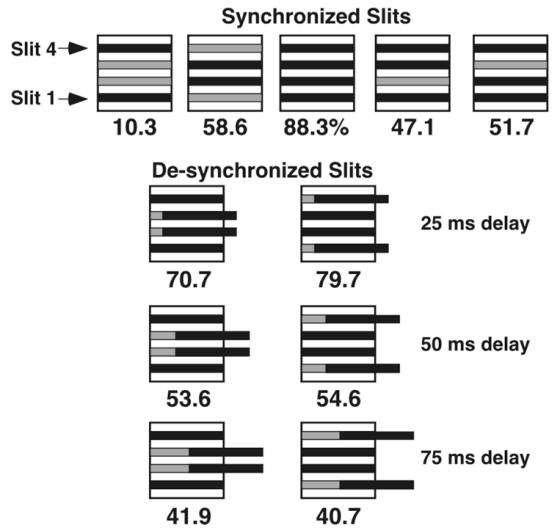


Fig. 7 Intelligibility (percent words correct) of sparse spectral sentences containing four narrow-band (1/3 octave) channels as a function of channel asynchrony. Note the relatively symmetrical decline in intelligibility associated with the central channels leading or lagging the lateral channels. 16 subjects. Adapted from [13].

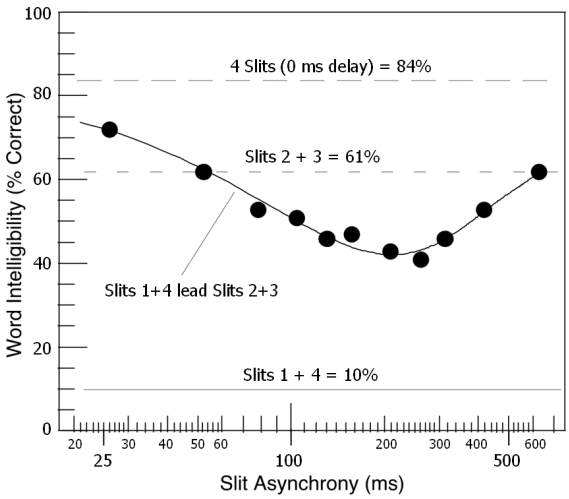


Fig. 8 Intelligibility of sparse spectral sentences containing four narrow-band (1/3 octave) channels as a function of slit asynchrony. Note that intelligibility goes below baseline (slits 2+3) when the slit asynchrony exceeds 50 ms). 27 subjects. Adapted from [21].

8. The Importance of Modulation Phase for Intelligibility

The results described in Sect. 7 imply that the relation between intelligibility and the low-frequency modulation spectrum is more complex than the formulation proposed by Houtgast and Steeneken using their Speech Transmission Index (STI) [18]. According to the STI, intelligibility is largely predictable from the contour and magnitude of the modulation spectrum. However, the results of Experiments 2 and 3 imply otherwise. If the modulation spectrum (as conventionally represented) was the primary determinant of intelligibility then each of the channels presented

in isolation would be highly comprehensible, as each is associated with a modulation spectrum that fits the conventional profile for intelligible speech (Fig. 9). Clearly, this is not the case, suggesting that some property of the acoustic signal *other* than the conventional modulation spectrum is essential for understanding spoken language. The results of Experiment 3 suggest that the phase of the modulation spectrum may be crucial for understanding spoken language. The temporal relation of the modulation patterns across the spectrum appear to be extremely important for decoding the speech signal. It is as if the underlying representation in the signal was derived from a complex topography of peaks and valleys distributed across frequency and time [14]. Both the phase and magnitude of the modulation spectrum would be required to reconstruct the topography with precision. If this were the case, then it should be possible to scramble the phase of the modulation spectrum using a different method and achieve a result comparable to that of Experiment 3.

In a fourth experiment this objective was accomplished using a full-spectrum version of the speech signal. Local time reversal of the signal's waveform, as shown in Fig. 10, provides a convenient means with which to dissociate the phase and magnitude components of the modulation spectrum. The experimental paradigm is loosely based on a study published by Saberi and Perrott [20]. Waveform segments of variable (and uniform) length were flipped on their horizontal axes and the intelligibility measured after such signal manipulations were imposed. As the length of the waveform interval increases, the amount of modulation-phase dispersion across the frequency spectrum increases substantially (relative to the original, undistorted signal), even though the impact on the magnitude component of the modulation spectrum is somewhat less dramatic. In fact, the magnitude of the modulation spectrum below 9 Hz actually increases for reversed waveform lengths of 50 ms or longer (Fig. 11).

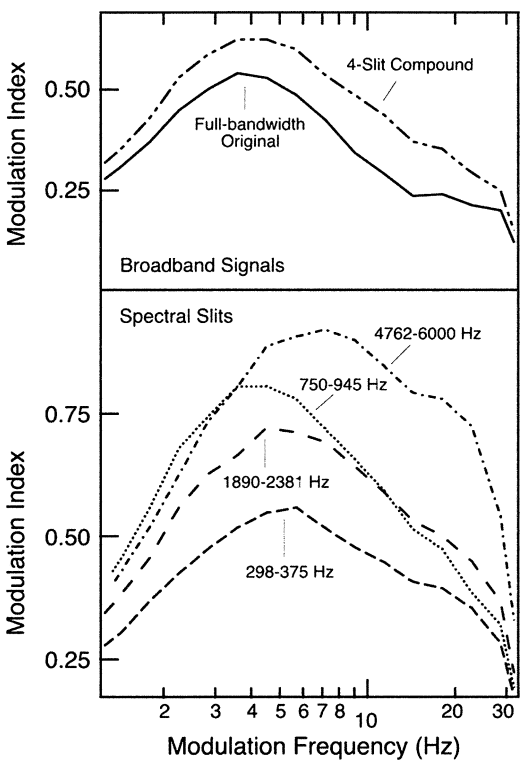


Fig. 9 The modulation spectrum (magnitude component) associated with each 1/3-octave channel, as computed for all 130 sentences presented in Experiment 2 [bottom panel]. The peak of the spectrum (in all but the highest channel) lies between 4 and 6 Hz. Its magnitude is considerably diminished in the lowest frequency slit. Also note the large amount of energy in the higher modulation frequencies associated with the highest frequency channel. The modulation spectra of the 4-channel compound and the original, unfiltered signal are illustrated for comparison [top panel]. From [13].

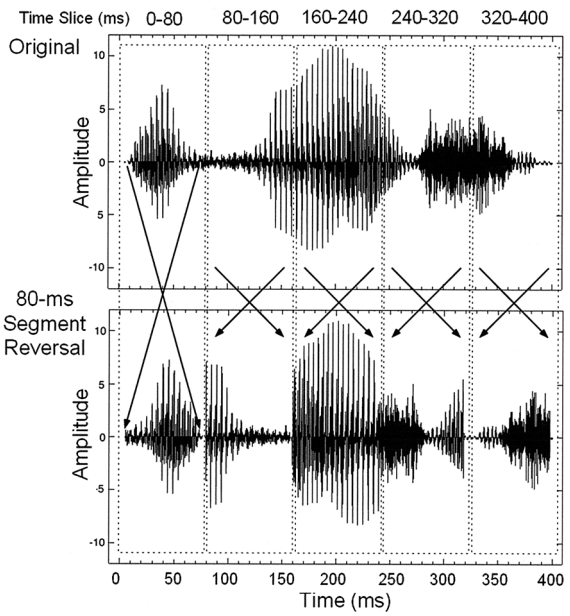


Fig. 10 The stimulus-processing procedure, as illustrated for a brief portion of a single sentence. Each segment was "flipped" on its horizontal axis, preserving all other temporal properties of the signal. In this example the reversed-segment duration is 80 ms. From [12].

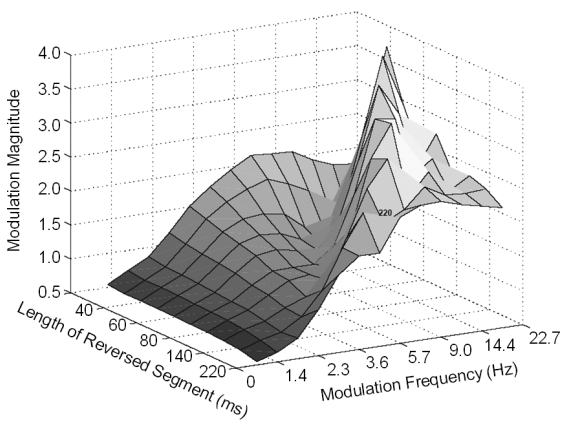


Fig. 11 The magnitude component of the modulation spectrum computed for all 40 sentences used in the experiment as a function of reversed-segment duration. Note that there is a slight decline in magnitude in the key, 3–9 Hz region for reversed-segment lengths of 20–50 ms, followed by a steep increase in magnitude for longer reversed-segment intervals. From [12].

When the phase and magnitude components are combined into a single representation (the "complex modulation spectrum" —[12]) the relation to intelligibility becomes much clearer. In this representation the magnitude of the modulation spectrum at each frequency depends on the phase coherence relative to the original signal. When modulation phase across the acoustic spectrum is coherent, the associated magnitudes sum linearly. Phase scrambling results in a diminution of complex modulation magnitude because the vectors associated with the magnitude component tend to cancel (Figure 12 illustrates the computation of the complex modulation spectrum).

The complex modulation spectrum progressively diminishes as the length of the waveform-reversal increases, and is closely paralleled by a concomitant decline in intelligibility (Fig. 13). Thus, the ability to understand spoken language appears to depend on *both* the magnitude and phase of the modulation spectrum distributed across the frequency spectrum. In some sense the fine acoustic-phonetic detail that figures so importantly in conventional accounts of the speech chain may merely reflect the modulation pattern distributed across the acoustic (tonotopic) frequency axis. This is hardly a novel concept, dating back to the development of the VOCODER in the late 1930's [6]. However, the importance of modulation seems to be have been largely forgotten over the intervening years.

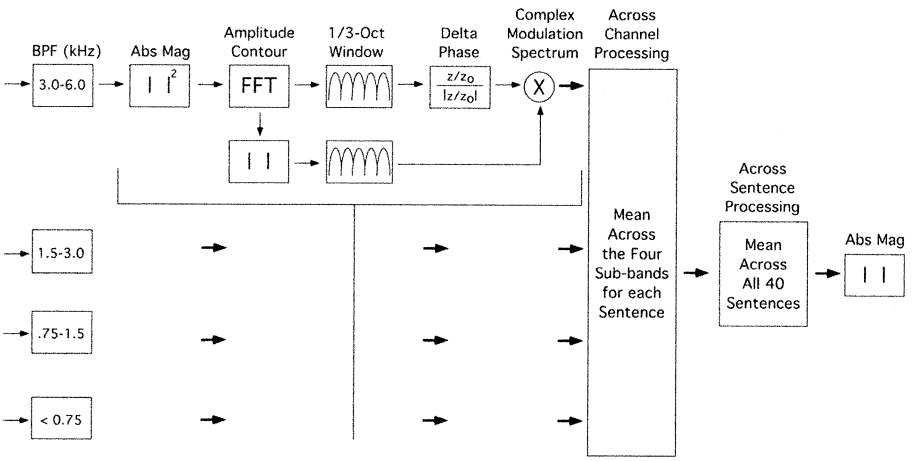


Fig. 12 Signal-processing procedure for computing the complex modulation spectrum of the sentential material. The magnitude and phase components of the modulation spectrum are initially computed separately for each of the four sub-bands. The delta-phase (relative to the original signal) is computed for 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 Fig. 13. From [12].

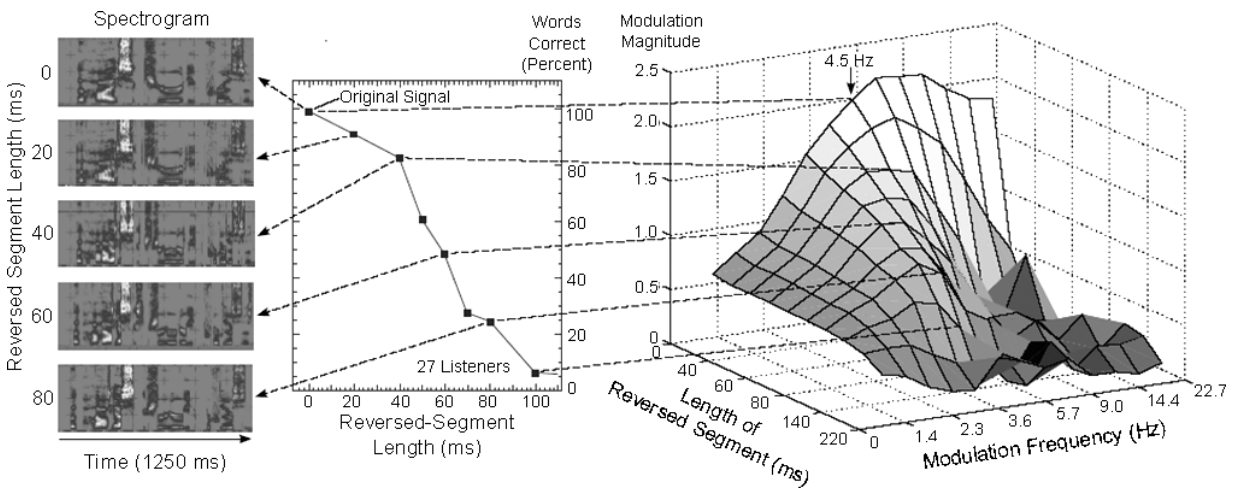


Fig. 13 Relation of intelligibility of locally time-reversed sentences (center) to the complex modulation spectrum (right) for reversed-segment durations between 20 and 100 ms (as well as the original signal). Spectrograms of a sample sentence are shown on the left. As the reversed-segment duration increases beyond 40 ms, intelligibility declines precipitously, as does the magnitude of the complex modulation spectrum. The spectro-temporal properties of the signal also degrade under such conditions. From [12].

9. The Role of Visual Information in Understanding Spoken Language

Despite the importance of the complex modulation spectrum for speech intelligibility, something else is missing from our formulation, namely the role played by the visual motion of the tongue, lips and jaw (as well as other facial features) in decoding spoken language. In this section we focus on an experiment that provides some insight into the underlying mechanisms that enable the brain to proceed from sound to meaning over a wide range of acoustic conditions using visual information to supplement that provided by the acoustic signal.

In this experiment the visual component of the speech signal (a talker producing short, simple sentences) was presented in tandem with a sparse spectral version of the same sentential material [9]. Only two acoustic channels were presented, one centered at 335 Hz, the other centered at 5400 Hz. The sentences were similar to (but distinct from) those used in Experiments 1–4. The two acoustic channels presented by themselves (A-alone) are associated with an intelligibility of ca. 20%. The visual information by itself (V-alone) results in only ca. 10% of the words being accurately reported. When the two modalities are combined (A-V) and presented synchronously, intelligibility rises to 63% overall for the nine listeners tested. The conditions described are associated with a 44–53% dynamic range in intelligibility. We can exploit the experimental paradigm to ascertain the temporal limits to A-V integration for decoding the speech signal by desynchronizing the audio and visual streams and then measuring this manipulation's effect on intelligibility. The impact of modality desynchronization is highly asymmetric and depends on which information stream (visual or audio) leads the other. When the audio signal is presented in advance of the video the intelligibility declines progressively with increasing amounts of asynchrony (Fig. 14). Except for a difference in baseline performance, the intelligibility function is similar to that of the acoustic-only condition described in Experiment 3 (in which the two central channels are desynchronized relative to their lateral counterparts). The performance functions run parallel to each other, suggesting that the mechanisms underlying the decoding of the speech signal under these two conditions are similar.

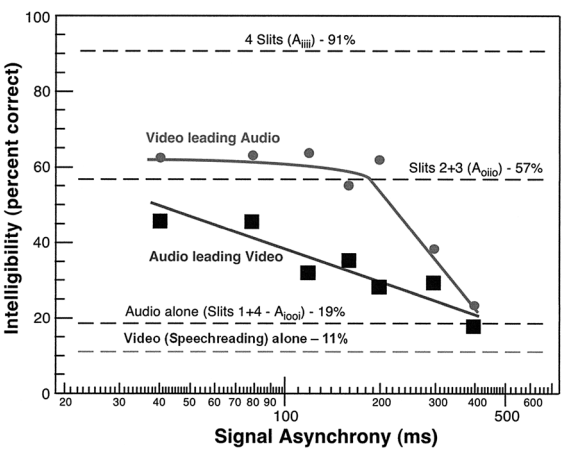


Fig. 14 Average intelligibility (for 9 subjects) associated with audio-visual speech recognition as a function of bi-modal signal asynchrony. From [9].

On the other hand, when the visual signal leads the audio, intelligibility remains unchanged overall until the two streams are desynchronized by more than 200 ms (Fig. 14). For larger amounts of asynchrony, intelligibility declines to baseline performance for intervals of 400 ms or longer. This is an intriguing result, as it implies that when the visual stream arrives first the brain is remarkably tolerant of asynchrony between acoustic and visual modalities (but not vice versa). What is the basis of this asymmetry? And can it provide additional insight into the mechanisms enabling the brain to decode spoken language?

A variety of potential "explanations" for this perceptual asymmetry come readily to mind.

A physical explanation, based on the disparate transmission speeds of acoustic and optical signals, does not account for the perceptual data observed. The speed of light is virtually instantaneous (299,792,458 m/s) while sound travels much more slowly, roughly one foot per millisecond (or 331.5 m/s). In the current experiment listeners were wearing headphones and viewing the visual stream on a monitor from a distance of a few feet. Thus, the physical time disparity between the arrival of audio and visual signals was a few milliseconds at most, nowhere close to the disparities observed in the experimental data.

Another prospective explanation pertains to the differential processing speeds of acoustic and visual information in the brain. If visual input is processed more slowly than the acoustic signal, then the early arrival of the visual signal could compensate for the differential processing speed, insuring that the two streams are internally processed in synch. Irrespective of the validity of this assumption (that visual information is processed more slowly than auditory), this hypothesis would imply that the intelligibility functions would be offset, but parallel to each other. However, the performance functions are of very different shapes, and therefore differential neural processing speed by itself, is unlikely to provide a full explanation for the perceptual asymmetry observed.

A third explanation combines elements of the first two hypotheses. Because the brain has evolved to process and decode information contained in sensory streams that are often of variable temporal relation to each other, it must have evolved mechanisms that are tolerant to A-V asynchrony. Although this hypothesis is appealing, it fails to account for the asymmetry observed in intelligibility. Some other factor must be at work. What might this be?

It has been shown by Grant and by others that much of the phonetic information contained in speechreading cues pertains to "place of articulation" — the characteristics distinguishing [p] from [t] and [k], or [b] from [d] and [g]

(see [8]). Such place cues reflect the locus of articulatory constriction during consonantal (particularly stop, fricative and nasal segments) production and are quite important for lexical discriminability (particularly at word onset). Such articulatory information is concentrated in the spectral region between 800 and 3000 Hz, and is often embedded in the dynamic patterns associated with formant transitions binding consonants and vowels. In this sense, place-of-articulation cues are inherently trans-segmental, spanning two or three phonetic segments in time [11].

Perhaps the perceptual asymmetry observed in Experiment 4 reflects "priming" properties of place-of-articulation cues contained in the visual stream? If this were so then one would anticipate that, under certain conditions, having the visual stream lead the audio would result in intelligibility that is even better than when the two streams are synchronized. This is precisely what happens for eight of the nine listeners in this experiment. Having knowledge of the visual signal in advance of the audio stream actually improves intelligibility slightly for asynchronies ranging between 80 and 120 ms. Why should this be so?

10. What is the Essence of Spoken Language?

Traditional models of spoken language focus on the phonetic segment (and its abstract representation, the phoneme) as the principal building block of words. However, the studies described in this paper suggest that the syllable is likely to play a far more important role than the phone in the lexical decoding process. The time intervals over which speech can be distorted and still be understood are more compatible with the syllable than the phone. Moreover, articulatory gestures associated with visible speech cues are syllabic rather than phonetic in nature. And yet the phone has figured importantly in models of speech for many decades. How can this be so if the phonetic segment plays a relatively minor role in lexical decoding?

Studies of pronunciation variation in spontaneous speech provide a potential explanation. Although words are commonly presented in terms of their phonetic constituents, oftentimes these phones are transformed in everyday utterances to something other than their canonical identity. The patterns of pronunciation variation provide important clues as to the underlying relation of the phone to the syllable.

The initial consonant(s) in a word (and syllable) are usually pronounced canonically (i.e., as represented in a dictionary) and such onsets are usually stable under a wide range of speaking styles and dialectal conditions.

In contrast, vocalic nuclei of the same words vary a lot. Such variation is conditioned by such factors as speaking style, geographical dialect, socio-economic status, emotionality and so on. Despite the enormous range of vocalic realization conditioned by such factors, certain principles apply that constrain the specific sorts of variation observed (at least in American English). Generally, pronunciation variants are of the same (or proximal) height as the canonical form, and it is rare that front and back variants are interchanged [14]. In other words, the basic "value" of a vowel remains preserved, even when its phonetic realization changes from a diphthong to a monophthong. The basic "place" within the articulatory vowel space remains relatively stable.

The final (coda) consonants lie somewhere in between the onset consonants and vocalic nuclei with respect to stability. In English, 75% of the coda consonants are coronals (mostly [t], [d] and [n]) [15] — such constituents manifest a propensity for "deletion" or "reduction" in spontaneous speech, particularly in unaccented syllables. This tendency is also common among syllable-final liquids ([l] and [r]) (though these constituents act more like vowels than consonants in many respects, and should therefore be considered as the terminal portions of the nucleus). Otherwise, coda consonants are relatively stable, like their onset counterparts.

The specific behavior of constituents with the syllable is largely conditioned by the syllable's prominence. Constituents within heavily accented syllables are usually canonically pronounced [17], providing some measure of support to the phonemic framework for spoken language. But such syllables are in the minority (at least for American English). The phonemic perspective appears to be, in part, a reflection of how words are supposed to be articulated under "ideal" (i.e., canonical) conditions. But there is more to the phonemic perspective than merely canonical pronunciation.

The three major articulatory dimensions for spoken language are voicing, place and manner of articulation. Manner of articulation (i.e., the distinction in the mode of production characterizing stops, nasals, fricatives, vowels, etc.) is temporally isomorphic with the concept of the phonetic segment [11]. The interval over which articulatory manner pertains also applies to the segment. Under virtually all conditions in which a pronunciation variant occurs there is stability in the inferred manner of articulation. For vocalic variants the manner is also preserved. In the case of coda deletions there is often the implication of a final constituent by virtue of the formant patterns remaining in the terminal portion of the vowel [15] as well as the fall in energy associated with that segment. In this sense, manner of articulation can be used to deduce the number (and often the identity) of the "underlying" segments, even when they are acoustically absent or exceedingly reduced. In this sense, the concept of the phone (and phoneme) is largely grounded in manner of production.

Within this framework, manner is coterminous with the segment, while voicing and place of articulation are not. Voicing is, according to this perspective, a syllabic feature, driven by prosodic factors, that can vary its temporal range across and within the conventional segment [11]. Thus, certain constituents are partially or entirely devoiced, while others exhibit more or less voicing than is typical of the canonical form. Place of articulation serves to bind the vocalic nucleus with onset and coda constituents, and in this sense is inherently trans-segmental [11]. It serves as the

articulatory basis of the di-phone and demi-syllable concepts used in speech synthesis. Place, like its articulatory counterparts manner and voicing, is sensitive to prosodic factors, and is particularly apparent in the visible speech cues used to decode the speech signal [8], [11].

The syllable serves as the organizational unit through which the articulatory dimensions of voicing, place and manner interact with each other and with prosodic information to shape the phonetic form of words within an utterance. Ultimately, such factors reflect the underlying information (both linguistic and paralinguistic) contained in the signal used by listeners to deduce the meaning associated with speech. The time constants associated with these processes are likely to be variable, ranging from ca. 10 ms for micro-phonetic phenomena to ca. 3000 ms for prosodic information. However, the core time interval of processing for speech intelligibility ranges between 40 ms and 400 ms, and largely reflects syllables and the articulatory constituents contained within.

11. Conclusions

The intelligibility of spoken language depends principally on low-frequency modulation of signal energy between 3 and 10 Hz. This low-frequency modulation reflects temporal properties of articulation associated with the syllable, and its phase characteristics are differentially distributed across the acoustic spectrum. This differential modulation phase characteristic is the acoustic basis for phonetic distinctions required for lexical discriminability, and is used in tandem with *visual* modulation information derived from the visible articulators, particularly in low signal-to-noise ratio conditions and among the hearing impaired. Prosodic properties of the linguistic signal are inherently linked to the modulation spectrum and provide a linguistically grounded framework with which to interpret the acoustic and visual components of the speech signal particularly under the sort of variable listening conditions characteristic of the real world.

References

[1] T. Arai and S. Greenberg, "The temporal properties of spoken Japanese are similar to those of English," Proc. 5th European Conf. Speech Commun. Technol. (Eurospeech-97), pp.1011–1014, 1997.

[2] T. Arai and S. Greenberg, "Speech intelligibility in the presence of cross-channel spectral asynchrony," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP-98), pp.933–936, 1998.

[3] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," J. Acoust. Soc. Am., vol.105, pp.2783–2791, 1999.

[4] M. Beckman, Stress and Non-Stress Accent, Fortis Publishing, Dordrecht, 1986.

[5] R. Drullman, J.M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am., vol.95, pp.1053–1064, 1994.

[6] H. Dudley, "Remaking speech," J. Acoust. Soc. Am., vol.11, pp.169–177, 1939.

[7] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP-92), pp.517–520, 1992.

[8] K.W. Grant, "Auditory supplements to speechreading," IEICE Technical Report, SP2003-43, 2003.

[9] K.W. Grant and S. Greenberg, "Speech intelligibility derived from asynchronous processing of auditory-visual information." Proc. Workshop Audio-Visual Speech Proc. (AVSP-2001), pp.132–137, 2001.

[10] S. Greenberg, "Speaking in shorthand: A syllable-centric perspective for understanding pronunciation variation," Speech Commun., vol.29, pp.159–176, 1999.

[11] S. Greenberg, "Pronunciation variation is key to understanding spoken language," Proc. 15th Int. Cong. Phon. Sci., pp.219–222, 2003.

[12] S. Greenberg and T. Arai, "The relation between speech intelligibility and the complex modulation spectrum," Proc. 7th European Conf. Speech Commun. Technol. (Eurospeech-2001), pp.473–476, 2001.

[13] S. Greenberg, T. Arai, and R. Silipo, "Speech intelligibility derived from exceedingly sparse spectral information," Proc. 5th Int. Conf. Spoken Lang. Process., pp.74–77, 1998.

[14] S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang, "Beyond the phoneme: A juncture-accent model for spoken language," Proc. 2nd Int. Conf. Human Lang. Technol. Res., pp.36–43, 2002.

[15] S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang, "The phonetic patterning of spontaneous American English discourse," Proc. ISCA/IEEE Workshop Spont. Speech Process. Recog., pp.35–38, 2003.

[16] S. Greenberg, S. Chang, and L. Hitchcock, "The relation between stress accent and vocalic identity in spontaneous American English discourse," Proc. ISCA Workshop on Prosody Speech Recog. Understanding, pp.51–56, 2001.

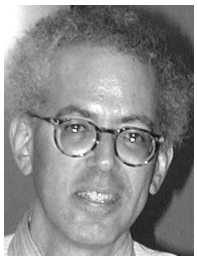
[17] L. Hitchcock and S. Greenberg, "Vowel height is intimately associated with stress accent in spontaneous American English discourse," Proc. 7th European Conf. Speech Commun. Technol. (Eurospeech-2001), pp.79–82, 2001.

[18] T. Houtgast and H. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am., vol.77, pp.1069–1077, 1985.

[19] W.D. Marslen-Wilson and P. Zwitserlood, "Accessing spoken words: The importance of word onsets," J. Exp. Psych.: Human Perception and Performance, vol.15, pp.576–585, 1989.

[20] K. Saberi and D. Perrot, "Cognitive restoration of reversed speech," Nature, vol.398, p.760, 1999.

[21] R. Silipo, S. Greenberg, and T. Arai, "Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations," Proc. 6th European Conf. Speech Commun. Technol. (Eurospeech-99), pp.2687–2690, 1999.



Steven Greenberg received an A.B. in Linguistics from the University of Pennsylvania (1974) and a Ph.D. in the same subject from the University of California, Los Angeles (1980). After completing his graduate studies he became a post-doctoral fellow, initially at Northwestern University (Audiology Program) and ultimately at the University of Wisconsin (Department of Neurophysiology). Dr. Greenberg remained at Wisconsin for nine years as a research scientist investigating the auditory bases of pitch perception and speech processing. In 1991 he moved to the University of California, Berkeley, where he established a speech research laboratory in the Department of Linguistics. In 1995 he moved to the International Computer Science Institute (affiliated with UC-Berkeley) in order to engage in basic research germane to automatic speech recognition. In 2002 he established the Speech Institute as a means of developing novel paradigms for speech technology research.

Takayuki Arai received the B.E., M.E. and Ph.D. degrees in electrical engineering from Sophia Univ., Tokyo, Japan, in 1989, 1991 and 1994, respectively. In 1992–1993 and 1995–1996, he was with Oregon Graduate Institute of Science and Technology (Portland, OR, USA). In 1997–1998, he was with International Computer Science Institute (Berkeley, CA, USA). He is currently Associate Professor of the Department of Electrical and Electronics Engineering, Sophia Univ. In 2003–2004, he is a visiting scientist at Massachusetts Institute of Technology (Cambridge, MA, USA). His research interests include signal processing, acoustics, speech and hearing sciences, and spoken language processing.