UNDERSTANDING SPEECH UNDERSTANDING: TOWARDS A UNIFIED THEORY OF SPEECH PERCEPTION

Steven Greenberg

University of California, Berkeley
International Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA
steveng@icsi.berkeley.edu

ABSTRACT

Ever since Helmholtz, the perceptual basis of speech has been associated with the energy distribution across frequency. However, there is now accumulating evidence that speech understanding does not require a detailed spectral portraiture of the signal. As a consequence, a new theoretical perspective, focused on time, is beginning to emerge. This framework emphasizes the temporal evolution of coarse spectral patterns as the primary carrier of information within the speech signal, and provides an efficient and effective means of shielding linguistic information against the potentially hostile forces of the natural soundscape, such as reverberation and background acoustic interference. The auditory system may extract this relational information through computation of the low-frequency modulation spectrum in the auditory cortex, and this representation provides a principled basis for segmentation of the speech signal into syllabic units. Because of the systematic relationship between the syllable and higher-level lexico-grammatical organization it is possible, in principle, to gain direct access to the lexicon and grammar through such an auditory analysis of speech.

1. INTRODUCTION

Although speech is the primary behavioral medium for human communication, the neurological and psychological processes underlying its perception are poorly understood. The traditional theoretical framework for speech has emphasized articulatory and cognitive aspects of its representation with comparatively little attention paid to auditory processes [48]. However, there is a growing consensus that these traditional paradigms are not sufficiently powerful to account for many of the most important perceptual properties of speech, and that a new conceptual framework is required.

1.1 The Origins of an Emerging Paradigm

The world in which we live differs dramatically from the one which arose from the ashes of the Second World War and gave shape to a form of scientific enterprise that profoundly influenced speech research. In those early years, the Sonograph™ was the analytical instrument of choice for speech research, by virtue of its detailed, seemingly objective spectro-temporal portraiture of the acoustic signal. The Sonograph appeared to offer a means of visualizing the perceptually relevant physical properties of the acoustic signal in sufficient detail as to provide a comprehensive technique for describing speech, and for correlating formant patterns with their underlying phonemic constituents. And because phonemes were thought to constitute the elementary building blocks of meaning, from which lexical units derive, it would be but a simple matter to trace the "speech chain," from sound to meaning.

However, as early as the late 1930's, research engineers at Bell Labs realized that such fine spectral detail was not required to satisfactorily reproduce speech. Dudley's channel vocoder quantized the spectrum into twenty or fewer channels, and passed through these filters only the low-energy fluctuations below 20 Hz to successfully resynthesize the signal [16].

In the 1950's a research group at Haskins Laboratories promoted yet a different means of sparsely representing the spectrum. Their spectral economization, known as the "pattern playback" [10] focused attention on the energy maxima ("formant patterns"). Because of the daunting challenge of encoding all of the possible co-articulatory effects imposed on these formant patterns, the Haskins group suggested that some more parsimonious representation of speech is likely to occur in the brain, one based on the underlying articulatory gestures generating the acoustic signal. Their proposal, now known as the "motor theory of speech perception" [40] has been highly controversial. The basic idea is intriguing. It appeals to the sensible intuition that similar mechanisms are likely to govern both the production and perception of speech and that some unifying representation must exist to enable a speaker to govern the acoustic output of the vocal tract with sufficient precision as to successfully manipulate the behavior of others through the force of spoken language. Despite its intuitive appeal, the motor theory has struck many as biologically far-fetched. It is difficult to fathom the nature of neurological mechanisms by which the brain could readily back-trace the articulatory gestures from the acoustic signal in real time.

But what if the brain were able to back-compute not the articulatory gestures, but rather the temporal dynamics that underlie both the production and acoustics of speech? And what if this information were recoverable from the appropriate auditory representation of the acoustic signal?

1.2 Representational Stability - the Essence of □□□□ Speech Understanding

It is the thesis of this paper that the temporal dynamics of the speech signal provide the key to understanding speech perception and afford a means of rendering the speech signal relatively impervious to the potentially deleterious effects of reverberation, background noise and speaker variation. A key issue for any theory of speech perception concerns the ability to create a functional equivalence across many diverse instances of the same basic meaning. How does the brain "know" that the acoustic patterns entering the ear at time x signify the same thing as the somewhat different patterns received at time x+d, where d can assume a value ranging from fractions of a second to years? This issue of perceptual "invariance" cuts to the very essence of speech perception. Any comprehensive theory must include a principled means of providing a stable representation across the full range of acoustic conditions typifying speech.

It is proposed that the auditory system captures the temporally dynamic properties of speech through computation of the low-frequency portion of the modulation spectrum and that this representation is remarkably stable over a wide range of acoustic environmental conditions. The low-frequency modulation spectrum, with it's emphasis on temporal intervals between 100 and 300 ms, is admirably suited to extract syllabic and related phonetic information required for accessing higher-level linguistic representations of the speech signal. Before discussing the modulation spectrum as a means of representing the speech signal, we briefly consider why the conventional spectral approach fails to successfully account for the perceptual

Proceedings of the ESCA Workshop on the "Auditory Basis of Speech Perception," Keele University, 1996, pp. 1-8

stability of speech and its robustness in the presence of acoustic interference.

1.3 The Ear as a Frequency Analyzer

The historical neglect of the auditory system as an explanatory basis for understanding the structure and function of speech is likely the consequence of viewing audition as a passive process. The auditory system's primary function was traditionally seen as computing a running spectrum of the acoustic signal for subsequent conversion into linguistic units by other parts of the brain [38]. Although auditory-based spectrograms might provide a more accurate means of visualizing the "internal" representation of the speech signal, they would not fundamentally alter the manner in which higher-level linguistic information is extracted from the acoustic signal. Thus, the role of the auditory pathway was viewed as largely confined to feeding spectra into a phonological processor which, in turn, churned phone sequences into a lexical units. The emergent linguistic properties of lexical reference, grammar and meaning were viewed as lying within the domain of higher cortical processes beyond the reach of the auditory pathway.

2. CRACKS IN THE SPECTRAL EDIFICE

One of the first indications that something was not quite right with this "bottom-up" view of speech perception was Warren's demonstration of "phonemic restoration." Complete occlusion of an entire phonetic constituent (e. g., the [s] in the word "legislature") by an interfering sound (e.g., a click or noise) is hardly noticed by listeners and has no apparent effect on intelligibility [60]. Speech understanding seemed to be anything but a data-driven, bottom-up process.

Moreover, Miller and Licklider's insightful study of the limits of intelligibility wrought by periodic (as well as random) interruption (and deletion) of the speech signal implied that much of the speech signal could be discarded without a significant impact on the decoding process [46]. At the time (1950), their findings were widely cited for demonstrating the "redundant" nature of speech. Phonetic features were said to be signaled by many different cues, distributed in both time and frequency, and this distributed representation had evolved to insure the robust transmission of information contained within the speech signal. But the specific nature of the representational distribution remained unspecified, and "redundancy" assumed the status of explanatory framework and became a common means of accounting for all sorts of curious and intriguing phenomena not readily accommodated within the strictly hierarchical view of speech perception.

2.1 Auditory Scene Analysis

A separate vein of scientific investigation, based on Gestalt principles of continuity and form, was initiated some 25 years ago by Al Bregman and his students. In the intervening years, Bregman [5], Darwin [11], McAdams [45] and others have elucidated many of the spectro-temporal constraints underlying the perceptual organization of speech and auditory function. However, their research provided no concrete biological foundation with which to bind these elegant perceptual demonstrations into a unified theoretical framework for understanding speech perception.

2.2 Computational Audition

Ten years ago this situation began to change. Weintraub, inspired by Marr's brilliant treatise on vision [43], applied computational methods to auditory scene analysis [62]. His work, building on Lyon's computational approach to physiologically plausible pitch extraction [42], set in motion a stream of ever more sophisticated computational research, culminating in the current crop of elegant models [e.g., 7, 9. 56] .

These computational models have enabled us to visualize how emergent perceptual properties pertaining to speech and other complex auditory phenomena, could arise from the activity patterns of simple neural elements. And though there has historically been considerable speculation on the roles played by frequency-selective and temporal mechanisms for the extraction of such perceptually relevant properties as pitch. loudness and timbre, the computational methods have provided a quantitative basis for visualizing such auditory representations.

2.3 Automatic Speech Recognition by Computer

An equally important development for evaluating the importance of auditory processes for speech understanding has stemmed from automatic (computer) speech recognition (ASR). Such systems attempt to recognize speech by building "word models" from sequences of phonetic segments ("phones") derived from abstract linguistic representations of speech called "phonemes." Proceeding from mainstream linguistic theory, each word is defined as a quasi-unique sequence of phonemes. Accurate decoding of a phoneme sequence should, in principle, provide the correct word most of the time. In those instances where the specific word is ambiguous, due to homophony and lexical neutralization, grammatical and syntactic context could be used to delimit the intended meaning.

The actual computational machinery underlying speaker-independent ASR is quite complex [49], involving highly sophisticated probabilistic (Hidden Markov) models (HMMs) for sorting through a combinatorially immense array of plausible alternatives given the acoustic evidence through the application of dynamic programming techniques utilizing the Viterbi algorithm. Some recent versions of ASR systems use a combination of HMMs and neural networks to fine tune the relation between the acoustic input and the candidate phonetic states [3].

What is common to virtually all current ASR systems is the reliance on essentially data-driven, bottom-up techniques for "boot-strapping" the recognition process. The speech signal is typically divided into 20-ms segments and each "frame" is associated with a vector of phonetic probabilities based on the computed similarity of that portion of the signal with a "composite" representation of a series of phone-like entities. It is at this stage that auditory principles have been allowed to intrude on the more conventional signal processing techniques traditionally used to characterize the speech frames.

Over the years it was discovered that the traditional spectral characterization of frames, based on the Fast Fourier Transform, was too detailed for adequate generalization of specific tokens (i.e., frames of speech) to the composite phonetic representations. Linear predictive coding (LPC), which provides a highly smoothed representation of the spectrum based on the inferred transfer function of the vocal tract producing the speech segment, was shown to improve the generalization, consistent with the articulatory perspective on speech decoding. However, refinements of the LPC representation, based on inferred auditory transformations of the input spectrum, have been demonstrated to achieve even better generalization. The two most popular auditory-inspired techniques are Mel-cepstrum [13] and Perceptual Linear Prediction (PLP) [30]. These techniques share in common an emphasis on the portion of the spectrum below 1500 Hz, commensurate with the spatial frequency organization of the human auditory system [28] and a highly smoothed spectral envelope. PLP incorporates further information pertaining to the audibility function of human hearing, spectral integration and a compressive loudness growth function to simulate the internal auditory representation of static spectra, refinements that have recently been shown to have a demonstrably positive effect on ASR performance. A more recent form of spectral conditioning, RASTA, which emphasizes the dynamic components of the speech signal, has also been shown to improve ASR performance under certain noisy conditions [31].

The traditional ASR approach performs well for limited vocabulary conditions such as spoken digits or numbers

(98-99% correct), or where the material consists of individuals speaking written sentences (e.g. TIMIT, Wall Street Journal achieve ca. 88 - 95% correct), and where the speech is recorded in a relatively noise-free acoustic environment. However, under many conditions simulating typical human communication, ASR systems perform much more poorly.

Computer speech recognition systems, capable of achieving high levels of performance on moderately complex linguistic material under pristine acoustic conditions, typically fail when confronted with comparable speech materials presented under more realistic listening conditions, in which reverberation and high levels of acoustic background interference commonly occur [23]. Addition of background noise or reverberation typically reduces the word-level accuracy to 20-50%. And even in the absence of such acoustic interference, naturally spoken discourse (e.g., the Switchboard corpus) with its attendant diversity of speaking styles and "sloppy" speech will humble even the most sophisticated speech recognition system [36].

What accounts for the disparity between the performance of ASR systems under ideal and more realistic conditions? Many practitioners of ASR believe that current deficiencies are correctable with a suitable expansion of speech material used to train the systems. However, increasing the training data is unlikely to substantially improve ASR performance because of the virtual impossibility of anticipating all of the acoustic conditions, speaker styles and pronunciation patterns likely to be encountered.

Human listeners do not re-calibrate the speech decoding process for every change in speaker, acoustic reverberation, background noise, rate of speaking, speaking, style, etc. and there is no reason, in principle, why machines should be required to do so either. A more complete and detailed understanding of how speech is processed by humans is likely to improve the performance of ASR systems under these more realistic conditions. But the current failure of ASR systems to perform well under real-world conditions can serve a useful scientific function, as it reveals the domains in which current models of speech perception are deficient. And ASR can also provide important clues as to the identity of organizational domains which have a significant impact on understanding spoken language.

2.4 Granularity of the Speech Spectrum

Under many circumstances only a coarse representation of the spectrum is required for accurate decoding of speech. At first glance this conclusion fails to conform to the auditory system's traditional role of frequency analyzer par excellence. The tuning of auditory neurons is relatively sharp and it would appear that the peripheral representation of the signal should provide an abundance of spectral detail for higher auditory centers to process for adequate phonetic classification. However, the sharp tuning of peripheral auditory neurons only pertains to low sound pressure levels, within ca. 30-40 dB of detection threshold (with the possible exception of the low-spontaneous-rate auditory-nerve fibers). Above this level the tuning broadens appreciably, particularly when measured in terms of neural synchrony [37]. Because speech generally is produced at relatively high sound pressure levels, typically only a few spectral features of the signal are actually encoded in the peripheral discharge patterns. These features are associated with spectral peaks, and for this reason computational techniques for spectral reduction, such as PLP are effective in capturing this dimensionality reduction in the representation of speech spectra.

Because of the speech signal's large range of acoustic variability (the result of heterogeneity in speaker vocal tracts, speaking style/rate and acoustic environmental conditions) a faithful, detailed representation of the speech spectrum would actually serve to impede effective generalization across the normal range of acoustic variation encountered. A potentially effective means of dealing with this overload of spectro-temporal detail is to encode only a sparse representation of the speech signal encapsulating the relevant linguistic information. Additional evidence in support of a sparse representation of the speech spectrum comes from cochlear implant patients who can often achieve a remarkably high degree of performance benefit, even without the aid of speech reading, from crude electrical stimulation patterns that provide little more than the low frequency modulation patterns distributed over just a few spectral channels [8]. And even in normal hearing individuals, the fine spectral detail of the speech signal is often blurred as a consequence of reverberant and noisy background conditions [47], further suggesting that listeners rarely encounter the archtypical spectral patterns found in speech texts.

If the full spectrum is not necessary for effective speech understanding, what is the minimum amount of spectral information required? Evaluations with a channel vocoder indicates that at least ten channels are required for a relatively faithful reproduction of speech. In informal tests Brian Kingsbury and I have found that the number of channels can be reduced to seven, if intelligibility rather than fidelity is the primary criterion of evaluation. But these estimates are based on voiced speech, with a clear source of glottal vibration. When the voicing source is quasi-white noise, effective intelligibility requires only three or four channels, as long as these bands partition the spectrum in a manner consistent with the range of the lower formants (Band 1 < 800 Hz, 800 Hz < Band 2 < 2500 Hz, Band 3 > 2500 Hz).

The physiological basis for this sparse spectral representation remains speculative, but probably relates to the ability of the auditory system to capture the temporal dynamics of the speech signal. The first format is generally restricted to frequencies below 800 Hz, the primary domain for neural phase-locking in the auditory periphery and central brainstem pathway. The third and higher formants typically occupy the region above 2500 Hz, where "place" mechanisms of spectral coding dominate. In between is the province of the second formant, a region where place and phase-locking mechanisms operate in tandem. Both Shannon's speech-modulated noise [55] and Ghitza's "tiling" [21] demonstrations partition the spectrum in a similar manner, with some measure of success. Allen [1], following Fletcher's Articulation Index (AI) framework [18], has suggested that the bandwidth of these correlated channels is roughly one octave, and that the activity pattern associated with such bands are essentially integrated into phonetic sub-features prior to their integration into more global representations.

2.5 Acoustic Shielding of Perceptually □□□ □□□□ Relevant Features

Neural phase-locking is thought to play an important role in the spectral processing of complex signals [52, 64]. To the extent that a detailed spectral representation of the speech signal is not required for adequate intelligibility, what might the role of this important medium of neural encoding be for the processing of speech?

One likely function is to shield the informational constituents of speech from the deleterious effects of background noise. Many auditory neurons preferentially discharge to low-frequency, quasi-periodic components of the acoustic signal, and this synchronized activity provides an effective means for the informationally relevant components of the signal to "rise above" the background. Noise, by virtue of its statistical properties, is not nearly as effective in capturing the temporal properties of such neurons [25].

The spectro-temporal properties of speech may have evolved as they have in order to provide a robust medium for transmitting information under variable acoustic conditions. The presence of nearly continuous voicing (i.e., pitch), the predominance of energy in the low-frequency (< 2 kHz) portion of the spectrum where neural phase-locking is

strongest, the prevalence of abrupt onsets for syllabic components (e.g., stop and affricate consonants) all suggest that a primary selection factor shaping the acoustics of speech was the ability to withstand the deleterious effects of the acoustic background [26] and to provide a means of grouping together the neural activity evoked by related spectral elements [9].

Binaural processing is another important mechanism for extracting speech-relevant information under noisy conditions. The auditory system appears to perform an operation analogous to a cross correlation in the representations of the signals reaching the opposing ears [63], and to effectively cancel much of the background noise as a result of this comparison process [58]. It is probably not coincidental that individuals with a substantial hearing loss in one ear often have difficulty understanding speech in noisy conditions, but not in quiet [6].

In view of the relative stability of linguistic information under conditions associated with significant acoustic variability of the speech signal, how is it possible for the brain to extract invariant representations?

2.6 Cross-spectral Integration of the Spectrum

Listeners appear capable of combining information across spectral regions to successfully decode the speech signal, and the integration's outcome can far exceed the sum of the analyses performed separately on the constituent bands. Both Warren [61] and Lippmann [41] have recently demonstrated that speech intelligibility based on spectrally delimited bands, separated by two or more octaves is far more accurate than would be expected on the basis of linear summation, posing a challenge for Fletcher's AI theory. Warren has also shown that even a narrow spectral band is capable of providing sufficient information to achieve ca. 35% correct in word-level decoding if strategically located ca. 1200-1500 Hz, the "pivot" region for inferring the movement of the second formant.

Such demonstrations suggest that the speech decoding process involves inferential tracking of the temporal dynamics over a few spectral regions, and although a detailed spectral representation provides the means to accomplish this objective, it is neither necessary, nor in certain circumstances, desirable. Preliminary efforts to apply such insights to ASR under noisy conditions show some measure of success [4].

3. TIME - THE UNIFYING DIMENSION

Although the mechanisms underlying this cross-channel integration are not well understood, the dimension of time is almost certainly involved. Local analysis of the spectrum appears to occur within milliseconds, while spectrally more global analyses often require tens to hundreds of milliseconds to perform.

3.1 Spectral Decorrelation Reveals Multiple Time Scales

This multiresolution time scale can be demonstrated by temporally decorrelating the output of critical-band channels through which speech has been passed. This decorrelation is accomplished by shifting the channel outputs in time relative to each other. If the integration of spectral information across channels occurs in "real time" then even a slight temporal shift in these channels would significantly decrease the intelligibility of speech. In fact, speech can withstand channel decorrelations as long as ca. 120 ms without significant loss in intelligibility, indicating the existence of at least two separate levels of analysis, one based on local, probably within critical-band information, and a second, based on global integration across channels. It is tempting to speculate that the local, "within-channel" auditory analyses reflect primarily the operation of peripheral and central brainstem processes, while across-channel operations, particularly those pertaining to correlation of features over multi-octave ranges involves cortical processes. The time course of these diverse operations will necessarily differ.

3.2 Interrupted Speech

How might we understand the auditory and higher cortical processes that permit the brain to reconstruct the linguistic message from only a small temporal portion of the original signal? Huggins [35] demonstrated that the intelligibility of interrupted speech crucially depends on two parameters - (1) a minimum acoustic duration of ca. 40 ms for individual segments and (2) an interval between successive segments of not more than 200 ms. As long as these two conditions are met, it is possible to introduce all sorts of temporal deletions and insertions without a significant decline in intelligibility. Huggins' results are in accord with the notion that intelligibility depends on the integrity of the low-frequency modulation spectrum. But what accounts for the 40- and 200-ms limits of these crucial time intervals?

3.3 Event Rates in the Auditory Nervous System

At the level of the auditory nerve and most auditory brainstem nuclei, the neural discharge rate is on the order of ca. 150-250 events per second. Even at the thalamic level, in the medial geniculate body, discharge rates of 100-200 spikes/s are not uncommon. However, in the cortex neuronal discharge rates rarely exceed 30 spikes/s, and more typically occur at rates between 5 and 20/s.

The reduced discharge rate in the auditory cortex likely reflects the preponderance of intra-cortical projections, which are themselves similarly constrained in their temporal resolution through other intra-cortical inputs. As a result, the auditory cortex is likely to function as a highly inertial system, akin to neural oscillators [7], in which thalamic input plays a relatively subordinate role except to signal major changes in the afferent input.

This reduction in discharge rate effectively "down-samples" the auditory representation, and as a result facilitates generalization across diverse instances of the "same" thing. The intra-cortical input is likely to contain information about the state of both adjacent and distant tonotopically organized elements, as well as about the temporal evolution of the spectrum.

It is likely that this across-channel analysis occurs in at least two stages, one associated with approximately. 40-ms intervals for phonetic, sub-feature analysis spanning several contiguous channels, and a longer (ca. 200-ms) interval required for integration of sub-featural information into a coherent representation for higher-level linguistic processing. These longer units, which correspond to roughly syllable-sized units, are distinguishable on the basis of the composition and order of these shorter sub-featural elements. Within this framework, phones can be thought as a constellation of sub-features which, when bound together across time, serve as the carrier of linguistic information through the action of the syllable, as exemplified in Ghitza's tiling experiments [21].

From whence do the 40- and 200-ms come? Are these time intervals specific to the auditory cortex? Or do they reflect a more general constraint on cortical processing independent of sensory modality?

3.4 Sensory-motor Integration

Two hundred milliseconds is a ubiquitous interval in measuring various aspects of sensory and motor function. The temporal integration epoch for both acoustic [17] and visual [50] stimuli is of this magnitude, as is the upper limit for the continuity effect in audition [32]. This interval is also about the minimum time for a motor reaction to occur, and appears to pertain to the integration time required for information emanating from the sensory portions of the brain to be placed in register with each other and with the motor system.

3.5 The Sensory Quantum

But elements within this 200-ms interval must also be distinguishable. This is where the 40-ms quantum interval plays an important role. There are many limits of auditory and visual sensation that conform to this length of time. The frame rate for motion pictures is 24 per second, a rate designed to insure the illusion of continuous motion. In audition, 40 ms is also the minimum segmental interval required for the illusion of continuity to occur [32]. It is also about the shortest span of segmented speech that can reliably be associated with a specific phonetic quality. In addition, this 40-ms interval corresponds to the interval in which acoustic stimulation begins to assume an independent identity. Darwin has shown that a spectral component which begins less than 40 ms prior to the beginning of a vocalic segment is not perceived as standing separate from the other components. When this same segment leads by more than 40 ms it is heard as a separate stimulus, with a clearly defined pitch, distinct from the vowel, although its presence contributes to the vocalic identity [12].

3.6 The Relation to Syllables and Phones

The typical length of a syllable in fluent speech is ca. 200 ms, [39] suggesting that this is the unit over which acoustic and visual information is integrated into a unitary perceptual and articulatory entity. 200 ms is sufficiently long as to provide some measure of perceptual stability through correlation with cortical activity across many parts of the brain, but is short enough to provide a sufficiently dynamic representation of the external stimulation as to be functionally effective. In some sense, we appear to interact with the external world at roughly five frames per second.

The auditory system appears to quantize syllabic units into ca. 40-ms frames and it is on basis of the composition of these shorter intervals that phonetic distinctions among syllables can be made. Auditory cortical recordings to syllables distinguishable on the basis of voice-onset-time are consistent with this idea [53].

4. THE IMPORTANCE OF SEGMENTATION

One of the most important roles played by the auditory system is to provide segmental information. It is known that the hearing impaired often gain significant benefit from speech reading [59], and this gain in intelligibility can be interpreted as the result of an independent source of information pertaining to speech segmentation and phonetic boundaries.

One of the paradoxes of hearing impairment is that the locus of energy in the speech signal (< 2 kHz) is considerably below that of the region showing the greatest deficit in sensitivity (typically > 3 kHz). How can this be if damage to the spectral analytic capability of the auditory system is the primary basis for the functional impairment caused by a sensori-neural hearing loss?

It is also known that the hearing impaired typically experience relatively little difficulty understanding speech in quiet, non-reverberant conditions [57], and that the single best predictor of speech intelligibility performance in quiet is the pure tone threshold for frequencies *below* 2 kHz [57]. However, under noisy conditions, the best predictor of speech intelligibility is the audiometric threshold *above* 2 kHz [57]. In other words, there appears to be something special about the mid and high-frequency regions that is extremely important for processing speech under noisy conditions, which is otherwise not so apparent. What might this function be?

A recent study indicates that the spectral region above 3 kHz is particularly important for delineating the segmentation and number of syllables in spoken language [24]. Anyone who has attempted to identify individual phonetic constituents of casually spoken speech can attest to the difficulty of the task in the absence of multi-syllabic context. Reliable information pertaining to syllabic segmentation appears to be essential for understanding spoken language [39] and syllabic structure is an important means of inferring the lexical identity of ambiguous speech [54]. Thus, in instances where the low-frequency portion of the spectrum is compromised by background noise it is likely that the higher-frequency portions of the speech signal assume additional significance. If segmental information associated with these channels is compromised, then the ability to understand spoken language will be impaired.

This evidence is consistent with the idea that the ability to understand speech relies as much on segmental analysis as it does on an analysis of the spectrum. In the absence of such segmentation, the ability to understand speech is severely compromised [15]. In its presence, comprehension occurs, even with minimal spectral cues [55]. This perspective is also consistent with repeated demonstrations of a significant gain in speech intelligibility under noisy conditions when visual information pertaining to the movement of the lips, jaw and facial musculature are combined with the acoustic signal for both normal and hearing-impaired individuals [59]. This visual analog of the speech signal is capable of providing important information pertaining to segmentation and to the phonetic identity of syllabic constituents by virtue of a temporal dynamic common to the acoustic and articulatory representations of speech.

5. THE SIGNIFICANCE OF THE SYLLABLE

The traditional unit of phonetic information is the phone, whose length in English typically runs between 50 and 150 ms (mean duration of ca. 100 ms). However, its segmentation in the speech signal is often difficult to specify with any degree of precision, in part because information pertaining to a phone often overlaps in time with information associated with adjacent phones. This phenomenon is often referred to as "co-articulation." Traditionally, the speech signal has been analyzed as a series of phones concatenated through co-articulation rules imposed by the biomechanical constraints of the vocal tract. It has been an article of faith that some component of the speech decoding process works back from the co-articulation to restore the individual phonemic constituents.

But what if the phone is not the basic unit of speech perception, and words are not represented in the brain as sequences of phonemes? What if the phone is actually a secondary unit of analysis whose major function is to distinguish among different forms of the basic perceptual/representational unit?

There is increasing evidence that this is indeed the case, and that the syllable, rather than the phone is the basic unit of speech perception.

Some of the evidence is indirect. The reaction times for identifying a target syllable is faster than for its constituent phones, even when the target phone is located at the beginning of the syllable [54]. An analysis of the literature on the effects of consonantal context on vocalic identity indicates that most, if not all of the effects reported, are a consequence of intra-syllabic segmentation. When most of a syllable is presented, vocalic identification is high. When significant portions of the syllable are missing, the intelligibility is much lower. Furthermore, most co-articulation effects occur within a syllable. Trans-syllabic co-articulation effects are comparatively small. Increases in speaking rate result in the deletion and mutation of most phonetic constituents - however, syllabic units are generally preserved. Articulations are generally programmed in syllabic, not phonemic units. Both speech-error mispronunciations and "tip-of-the-tongue" recalls are organized on the basis of syllabic, not phonemic entities [20]. And integration of visual cues in speech perception occurs over syllabic, not phonemic intervals [44].

Discarding the phone in favor of the syllable would go a long way towards mitigating, if not eliminating many of the theoretical difficulties that have long plagued speech

research. And it would provide a means for systematically incorporating prosodic properties such as pitch, accent and stress, that have been difficult to achieve within the traditional phonological framework since these phenomena are organized on the syllabic, rather than on the phonemic level.

6. THE MODULATION SPECTRUM

Nearly two decades ago Plomp, Houtgast and associates began to investigate the importance of low-frequency modulations for the encoding of speech information. Their basic idea is that phonetic information can be encoded in terms of the slow energy fluctuations that occur across tonotopically organized auditory channels. Long-term analysis of the energy fluctuations indicates a peak in the spectrum at around 4 Hz [33], corresponding to the rate of syllabic units. And though the granularity of the spectral information would necessarily be coarse, it would be sufficient to adequately distinguish among the possible set of phonetic elements. However, the general implications for theories of speech perception in general were not widely appreciated, despite a prescient paper by Haggard [29].

Recently, Drullman and colleagues have demonstrated that the intelligibility of Dutch words and sentences is dependent on the integrity of the low-frequency portion of the modulation spectrum [15]. Low pass filtering the modulation spectrum, so that energy fluctuations above 3-4 Hz are significantly attenuated, reduces the intelligibility of speech, a finding that has been replicated for both English and Japanese [2]. This form of filtering leaves the quasi-steady-state spectral regions relatively intact, but essentially blurs the syllabic boundaries.

Brian Kingsbury, Nelson Morgan and I have recently developed a means of visualizing speech in terms of these low-frequency modulation characteristics to ascertain if this representational form remains stable under conditions which are known to preserve intelligibility, but disrupt more traditional representations based on spectrographic analysis. The representations are modulation spectrograms which encode the magnitude of energy in the lowest modulation band (with a peak at 4 Hz and 10 dB down at 8 Hz, similar to the long-term modulation spectrum of continuous speech) as a function of frequency (quantized to critical-band like units) and time (using 250-ms windows, and 12.5 ms steps to capture the dynamic aspects).

The traditional spectrographic display undergoes dramatic degradation in the presence of background noise and reverberation, under conditions which have little impact on speech intelligibility. The modulation spectrograms are remarkably stable under the same conditions, suggesting that information in the low-frequency modulations may be sufficient to encode speech-relevant information. (the spectrograms are visualizable via a WWW site, http://www.icsi.berkeley.edu/~steveng/modspec). It is also of interest that the modulation spectrograms are similar, in many respects, to the population response of auditory cortical neurons, which are most responsive to modulation frequencies lower than 20 Hz [52] and may therefore approximate the representation of speech signals in at least some proportion of cells in the human auditory cortex.

7. FROM SOUND TO MEANING

A major challenge for speech science is to specify the processes by which the brain is able to use the auditory modality for accessing lexical and semantic information. The traditional framework requires a complex and rather arbitrary series of operations that proceed from phonemic units to words, and from words to meaning via a language's grammar. However, even a cursory examination of the broad statistical properties of speech indicates that the relation between sound and symbol (at the lexical and grammatical elements) is anything but arbitrary. Such acoustic-linguistic associations could be utilized by the brain to fashion meaning from the acoustic stream.

Although only a quarter of the possible words in the English lexicon are one syllable in length, over 82% of the words in spoken discourse are mono-syllabic [19, 27]. What accounts for this large disparity between the dictionary and observed English usage?

Zipf observed many years ago that for *written* language the length of a word is inversely correlated with its frequency [65]. Mandelbrot reformulated Zipf's law in terms of information, and we may extend this insight to account for the preponderance of short, monosyllabic word in spoken discourse.

The reaction time for word recognition (presented visually) is known to be inversely proportional to its frequency of occurrence [34]. From this observation it is logical to surmise that there is a direct relationship between a word's information content and its duration. The higher the information content of a lexical item (i.e., the less predictable it is) the longer in duration it is likely to be. This relationship makes sense on the assumption that the temporal properties of language are tailored to synchronize retrieval from lexical memory with the time course of individual verbal elements. On this hypothesis, high-frequency words tend to be short because of their predictability and the relatively short time to retrieve their meanings, while rare words require longer retrieval time and are therefore often contain many syllables. Over the course of a word's linguistic evolution, it will tend to shorten as its usage increases with familiarity (e.g., "automobile" > "auto," " car"; "airplane" > "plane"; "refridgerator" > "fridge").

Many of the short, mono-syllabic words are so-called "function" words, such as articles (e.g., "a," "the"), prepositions (e.g., "of," "in"), conjunctions (e.g., "and," "or"), pronouns (e.g., "I," "you") and auxiliary verbs (e.g., "have," "would") which form the linguistic scaffold for spoken discourse. Of the one hundred most frequently spoken words, all but three are one syllable in length [14], and most of these are function words.

Poly-syllabic words, particularly those of three syllables or longer tend to be nouns or nominal modifiers, such as adjectives. It is as if the real-time demands of speech production and perception do not allow for the luxury of many "high-cost" words. Better to get one's point across with simple words than to take a chance with more formal, elegant means of expression.

In written language there is still a preponderance of mono-syllabic words (63%), but it is clear that the greater time which writing and reading affords, allows the writer to use more elegant and specific words than spoken language allows.

Most contemporary writing systems are based on alphabetic (i. e., phone-based) systems for reasons of efficiency and economy of expression, not for accuracy of reproducing spoken words. The earliest orthographic systems were based on either word (early Sumerian cuneiform) or syllable (later Sumerian cuneiform, Mycenean Linear B) units. The Chinese writing system currently in use is essentially logographic-syllabic in format (as a consequence of Chinese lexemes being monosyllabic). The mismatch between orthographic convention and spoken linguistic representation may very well underlie the difficulty with which many young children learn to read using alphabetic systems. Syllable-based intermediaries have been demonstrated to serve as effective pedagogical tools for children who have not been able to read using the traditional phonics approach [22].

The syllable has often been dismissed as a likely candidate for representing lexical information in English by virtue of its potentially complex and heterogeneous nature. In many languages (e.g., Japanese, Spanish) syllables are composed of a sequence of alternating consonants and vowels (e.g., CV, CVC, VC), while English contains syllables of the form CCCVCC (e. g., "strength") and

CCVCC (e.g., "cracked"), as well as the more transparent forms of the alternating consonant-vowel variety.

The division between written and spoken expressions of linguistic material have led many to erroneously conclude that the syllable is not likely to be a primary unit of speech perception and representation in English. In spoken discourse, over 80% of the syllables are of the canonical CV, CVC, VC, V form, and many of the remainder reduce to this format by processes of assimilation and reduction. The remaining exceptions are themselves linguistically marked by virtue of this deviation from the archetype, and tend to be either low-frequency nouns or inflected verbs (e.g., "look" [CVC] versus "looked" [CVCC], "looks" [CVCC]). In either instance, a deviation from the expected syllabic form provides important information, potentially useful for inferring its lexical and grammatical status.

In many languages of the world, with more transparent syllabic structure (i.e., of the CV, CVC variety) grammatical markings are typically imposed through affixing (i. e., concatenation of syllables) rather than through complexification of the syllable. These "agglutinative" languages (e.g., Turkish) stand in opposition to the more "synthetic" languages (e.g., Salish) by virtue of making necessary grammatical and semantic distinctions through the serialization of simple syllabic entities, rather than modifying the syllabic root. Some languages use a combination of the agglutinative and synthetic strategies. Regardless of the specific strateg(ies) adopted for encoding and representing this important higher-level linguistic information, the syllable stands at the nexus between sound and meaning.

Although significant differences separate the written and spoken forms of a language, in certain respects they are remarkably similar in terms of their statistical properties. In both instances, there is a reliance on a core vocabulary for expression of semantic information. In written English just 9 words form 25% of the total words used, 69 words account for 50% of word usage and 732 words account for 75% of lexical instances [14]. The corresponding statistics for spoken English show an even greater reliance on a core vocabulary [ ].

There is a similar reliance on a core body of elements at the syllabic level. For written English, 12 syllables form over 25% of all syllables used, 70 syllables constitute over 50% of syllabic occurrences and 339 syllables account for 75% of syllabic usage [14]. A similar pattern obtains for spoken discourse [19, 27]. An often cited disadvantage of a syllabic representation of English is the large number of distinct units required to cover the lexical inventory. But in actual practice the number of commonly used syllables is relatively small. And the remainder can be derived from relatively simple phonetic extensions to the core syllabic inventory. In many languages the issue of syllable inventory does not even arise. For in languages, such as Japanese, with a relatively transparent syllabic structure, the total number of separate syllables does not exceed a few hundred.

Listeners are capable of understanding spoken language by virtue of perceptual strategies that appear to automatically extract syllable-like units in the speech stream through analysis of the low-frequency modulation spectrum of the acoustic signal. Because of the systematic relationship between a language's syllable structure and its higher level semantics and grammar, and through a reliance on a core vocabulary of a few hundred, highly familiar lexical items, the brain is able to derive meaning from the speech signal on a continuous basis.

ACKNOWLEDGMENTS

I would like to thank Su-Lin Wu and Joy Hollenbach for computing some of the statistical analyses of English, as well as members of the ICSI Realization Group for valuable discussion on many issues described in this paper. I am also grateful to Mike Shire for assistance with the time-shifted speech demonstrations and Brian Kingsbury for computing the modulation spectrograms.

The support of the National Science Foundation, U. S. Department of Defense, and the European Community is gratefully acknowledged.

REFERENCES

[1] Allen, J.B. (1994) How do humans process and recognize speech?. *IEEE Transactions on Speech and Audio Processing*, 2, 567-577.

[2] Arai, T., Hermansky, H. Pavel, M. and Avendado, C. (1996) Intelligibility of speech with high-pass filtered time trajectories of spectral envelopes. Proc. ICSLP.

[3] Bourlard, H. A. and Morgan, N. (1993) *Connectionist Speech Recognition: A Hybrid Approach*. Boston: Kluwer.

[4] Bourlard, H., Dupont, S. and Morgan, N. (1996) A new ASR approach based on independent processing and recombination of partial frequency bands. *ICSLP Proc.*

[5] Bregman, A. S. (1990) *Auditory Scene Analysis*. Cambridge: MIT Press.

[6] Bronkhorst A. W. and Plomp R. (1989) Binaural speech intelligibility in noise for hearing-impaired listeners. *J. Acoust. Soc. Am.* 86, 1374-1383.

[7] Brown, G. and Cooke, M. (1996) Are neural oscillators the substrate of auditory grouping? Proc. ESCA ETRW Auditory Basis of Speech Perception (this volume).

[8] Clark, G. M. (1994) The development of speech processing strategies for the University of Melbourne/Cochlear multiple channel implantable hearing prosthesis. Special Issue: Cochlear implants. *J. of Speech-Lang. Path. Audiol.*, 16, 95-107.

[9] Cooke, M. (1993) *Modeling Auditory Processing and Organization*. Cambridge: Cambridge University Press.

[10] Cooper, F. C., Delattre, F P. C., Liberman, A. M., Borst, J. M., Gerstman, L. (1952) Some experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Am.*, 24, 597-606.

[11] Darwin, C. J. and Carlyon, R. P. (1995) Auditory grouping, i in *Hearing. Handbook of Perception and Cognition*, 2nd ed., B. C. J. Moore (Ed.), San Diego: Academic Press, pp. 387-424.

[12] Darwin, C. J.; Sutherland, N. S. (1984) Grouping frequency components of vowels: When is a harmonic not a harmonic? *Quart. J. of Exp. Psych. (HMM)*, 36, 193-208.

[13] Davis, S.B.; Mermelstein, P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Proc.*, ASSP-28, 357-366.

[14] Dewey, G. (1923) *Relative Frequency of English Speech Sounds*. Cambridge: Harvard University Press.

[15] Drullman R; Festen J. M. and Plomp, R. (1994) Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, 95, 1053-1064.

[16] Dudley, H. (1939) Remaking apeech. *J. Acoust. Soc. Am.*, 11, 169-177.

[17] Eddins, D. A. and Green, D. M. (1995) Temporal integration and temporal resolution, in *Hearing. Handbook of Perception and Cognition*,, 2nd ed., B. C. J. Moore (Ed.), San Diego: Academic Press, pp. 207-242.

[18] Fletcher, H. (1953), *Speech and Hearing in Communication*, Princeton: van Nostrand.

[19] French, N. R., Carter, C. W. and Koenig, W. (1930) The words and sounds of telephone conversations. *Bell System Tech. J.*, 9, 290-324.

[20] Fromkin, V. A. (1973) Slips of the tongue. *Sci. Am.*, 229, 110-117.

[21] Ghitza, O. (1993) Processing of spoken CVCs in the auditory periphery: I. Psychophysics. *J. Acoust. Soc. Am.*, 94, 2507-2516.

[22] Gleitman, L. R. and Rozin, P. (1973) Teaching reading by use of a syllabary. *Reading Research Quarterly*, 8, 447-483.
[23] Gong, Y. (1995) Speech recognition in noisy environments: A survey. *Speech Communication,* 16, 261-291.
[24] Grant, K. W. and Walden, B. E. (1996) Spectral distribution of prosodic information. *J. Speech Hearing Res.*, 39, 228-238.
[25] Greenberg, S. (1988) The ear as a speech analyzer, *J. Phon .*, 16, 139-150.
[26] Greenberg, S. (1995) The ears have it: The auditory basis of speech perception, in the *Proc. of the ICPhS*, 3, 34-41.
[27] Greenberg, S. and Wu, S.-L. (1996) Properties of English syllables in a spoken language corpus (Switchboard), unpublished data. Int. Comp. Sci. Inst.
[28] Greenwood, D.D. (1961) Critical bandwidth and the frequency cooedinates of the basilar membrane. *J. Acoust. Soc. Am.* 33, 1344-1356.
[29] Haggard, M. (1985) Temporal patterning in speech: The implications of temporal resolution and signal processing, in *Time Resolution in Auditory Systems*, A. Michelson (ed.). Berlin: Springer, pp. 217-237.
[30] Hermansky, H. (1990) Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.*, 87, 1738-1752.
[31] Hermansky, H. and Morgan, N. (1994) RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2, 578-589.
[32] Houtgast, T. (1974) *Lateral Suppression in Hearing*. Thesis. University of Amsterdam.
[33] Houtgast, T and Steeneken, H. (1985) 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.
[34] Howes, D. (1967) Equilibirum theory of word frequency distributions. *Psychnom. Bull.* 1, 18.
[35] Huggins, A. W. (1975) Temporally segmented speech. *Percept. Psychophys.*, 18, 149-157.
[36] Jeanrenaud, P., Eide, E., Chaudhari, U., McDonough, J., Ng, K, Siu, M. and Gish, H. (1995) Reducing word error on conversational speech from the switchboard corpus. *IEEE ICASSP*, 1, 53-56.
[37] Jenison, R., Greenberg, S. and Kluender, K. (1991) A composite model of the auditory periphery for the processing of speech based on the filter response functions of single auditory nerve fibers. *J. Acoust. Soc. Am.* 90, 773-786.
[38] Klatt, D. H. (1989) Review of selected models of speech perception. in *Lexical Representation and Process.*, W. Marslen-Wilson (ed), Cambridge: MIT Press, pp. 169-226.
[39] Lehiste, I. (1970) *Suprasegmentals.* Cambridge: MIT Press..
[40] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., Studdert-Kennedy, M. (1967) Perception of the speech code. *Psych. Rev.*, 74, 431-461.
[41] Lippmann, R. P. (1996) Accurate consonant perception without mid-frequency speech energy. *IEEE Trans. Sp. Aud. Proc.*, 4, 66-69.
[42] Lyon, R. F. (1984) Computational models of neural auditory processing.*IEEE ICASSP* , 3611-3614.
[43] Marr, D. (1982) *Vision*. San Francisco: W. H. Freeman.
[44] Massaro, D. W., Cohen, M. M. (1993) Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication*, 13, 127-134.
[45] McAdams, S. (1993) Recognition of sound sources and events, in *Thinking in Sound: The Cognitive Psychology of Human Audition*. S. McAdams and E. Bigand (Eds.) Oxford: Oxford University Press, pp. 146-198.
[46] Miller, G. A. and Licklider, J. C. (1950) The intelligibilty of interupted speech. *J. Acoust. Soc. Am.*, 22, 167-173.
[47] Payton, K. L., Uchanski, R. M. and Braida, L. D. (1994) Intelligibility of conversational and clear speech in noise and reverberation∂ for listeners with normal and impaired hearing. *J. Acoust. Soc. Am.*, 95, 1581-1592.
[48] Pisoni, D. B. (1982) Speech perception: The human listener as cognitive interface. *Speech Technology* , 1, 10-23.
[49] Rabiner, L. R. and Juang, B.-H. (1993) *Fundamentals of Speech Recognition*. Englewood Cliffs, N.J.: Prentice Hall.
[50] Regan, D. and Tyler, C. W. (1971) Temporal summation and its limit for wavelength changes: An analog of Bloch's law for color vision. *J. Optic. Soc. Am.,* 61, 1414-1421.
[51] Schreiner, C. E. and Langner, G. (1988) Coding of temporal patterns in the central auditory nervous system, in *Auditory Function: Neurobiological Bases of Hearing*, G. M. Edelman, W. E. Gall and W. M. Cowan (eds.). New York: Wiley, pp. 337-361.
[52] Schreiner, C. E. and Urbas, J. V. (1986) Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). *Hearing Res.* 21, 227-241.
[53] Schreiner, C. and Wong, S. (1996) Spatial-temporal representation of syllables in cat primary auditory cortex. Proc. ESCA ETRW Auditory Basis of Speech Perception (this volume).
[54] Segui, J. Dupoux, E. Mehler, J. (1990) The role of the syllable in speech segmentation, phoneme identification, and lexical access, in *Cognitive Models of Speech Processing: Psycholinguistic and computational perspectives*. G. Altmann, (Ed.) Cambridge: MIT Press, pp. 263-280.
[55] Shannon, R. V., Zeng, F.-G., Kamath, V.and Wygonski, J. Speech recognition with primarily temporal cues. *Science*, 270, 303-304.
[56] Slaney, M. and Lyon, R. F. (1993) On the importance of time - a temporal representation of sound, in *Visual Representations of Speech Signals*. Cooke, M., Beet, S. and Crawford, M. (eds.). Chichester: Wiley, pp. 95-116.
[57] Smoorenburg, G. F. (1992) Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *J. Acoust. Soc. Am.*, 91, 421-437.
[58] Stern, R. M. and Trahiotis, C. (1995) Models of binaural interaction, in *Hearing. Handbook of Perception and Cognition*, 2nd ed., B. C. J. Moore (Ed.), San Diego: Academic Press, pp. 347-386.
[59] Summerfield, Q. (1992) Lipreading and audio-visual speech perception, in *Processing the Facial Image*, V. Bruce, A. Cowey, A. W. Ellis, D. I. Perrett (Eds.). Oxford: Oxford University Press, pp. 71-78.
[60] Warren, R. M. and Obusek, C. J. (1971) Speech perception and phonemic restorations. *Percept. Psychophys.*, 9, 358-362.
[61] Warren, R. M., Riener, K. R., Bashford, J. A. and Brubaker, B. S. (1995) Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Percept. Psychophys..*, 57, 175-182.
[62] Weintraub, M. (1985) *A Theory and Computational Model of Monaural Auditry Sound Segregation*, Ph.D. Thesis. Stanford University..
[63] Yin, T. C. T. and Kuwada, S. (1984) Neuronal mechanisms of binaural interaction, in Edelman, G. M., Gall, W. E., Cowan, W. M. (eds.), *Dynamic Aspects of Neocortical Function.*, New York: Wiley, pp 263-313
[64] Young, E. D. and Sachs, M. B. (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of auditory-nerve fibers. *J. Acoust. Soc. Am* .66, 1381-1403
[65] Zipf, G. K. (1945) The meaning-frequency relationship of words. *J. Gen. Psych.*, 33, 251-256.