

THE EARS HAVE IT: THE AUDITORY BASIS OF SPEECH PERCEPTION

*Steven Greenberg
Department of Linguistics
International Computer Science Institute
University of California, Berkeley, CA 94720 USA*

ABSTRACT

Two neural systems - the auditory pathway and the thalamo-cortical association areas - set constraints on the acoustic properties of all vocal communication systems, including human language. This paper discusses the manner in which the two systems, operating in concert, may shape the spectro-temporal properties of human speech.

INTRODUCTION

The acoustic nature of speech is typically attributed primarily to constraints imposed by the human vocal apparatus. The tongue, lips and jaw can move only so fast and so far per unit time, the size and shape of the vocal tract set limits on the range of realizable spectral configurations, and so on. Although such articulatory properties doubtless impose significant constraints, it is unlikely that such factors, in and of themselves, entirely account for the full constellation of spectro-temporal properties of speech. For example, there are sounds which the vocal apparatus can produce, such as coughing and spitting, which do not occur in any language's sound inventory. And although speech can readily be whispered, it is only occasionally done. The vocal tract is capable of chaining long sequences of vowels or consonants but no language relies exclusively on either basic segment form, nor does speech contain long sequences of acoustically similar segments.

In this paper it is proposed that auditory system imposes its own set of constraints on the acoustic nature of speech, and that these factors are crucial for understanding how information is packaged in the speech waveform. This

packaging is designed largely to insure robust and reliable coding of phonetic information by auditory mechanisms operating under a wide range of potentially adverse acoustic conditions, as well as to integrate these phonetic features with information derived from neural centers responsible for visual and motor coordination.

SPECTRO-TEMPORAL PROPERTIES OF SPEECH

The discussion focuses on the following parameters of auditory function:

- (1) the range of frequency sensitivity
- (2) the frequency resolving capabilities of peripheral auditory neurons
- (3) the limits of neural periodicity coding
- (4) the time course of rapid adaptation
- (5) the temporal limits of neural coincidence detection
- (6) the modulation transfer characteristics of brainstem and cortical auditory neurons.

These parameters account for a number of important acoustic properties of speech, including:

- (1) an absence of spectral energy above 10 kHz
- (2) a concentration of spectral information below 2.5 kHz
- (3) preference for encoding perceptually relevant information in the spectral peaks
- (4) sound pressure levels for segments ranging between 40 and 75 dB
- (5) rapidly changing spectra
- (6) a prevalence of phonetic segments with abrupt onsets and transients
- (7) a preference for quasi-periodic waveforms whose fundamental frequencies range between 75 and 330 Hz

- (8) temporal intervals for integrative units (syllables and segments) ranging between 50 and 250 ms

These acoustic properties are essential for the robust encoding of speech under uncertain (and often noisy) acoustic conditions and in circumstances where multiple vocal interactions may occur (e. g., the proverbial cocktail party).

Each of these properties is examined in turn.

Spectral Energy Distribution

- a. Energy entirely below 10 kHz.

No speech segment contains significant energy greater than 10 kHz, and most speech segments contain little energy above 4 kHz [1,2].

- b. Concentration of energy below 2.5 kHz.

The long term power spectrum of the speech signal is concentrated below 2.5 kHz, as a consequence of the temporal domination of vowels, glides and liquids in the speech stream [1,2].

Moderate to High Sound Pressure Levels

Although we are certainly capable of talking at very high or very low amplitudes, we rarely do so. Rather, speech is spoken at a level that is approximately 60-75 dB for vocalic segments and about 20-30 dB lower for stops and fricatives [3].

Rapid Changes in Spectral Energy over Time

- a. Significant changes in spectral energy maxima over 100-ms intervals.

The spectral energy in speech moves rapidly through time. There is virtually no segment in the speech signal where the formant pattern remains relatively stationary for more than 50 ms [2].

- b. Prevalence of abrupt onsets (e. g. stop consonants, clicks, affricates).

There is a tendency for words to begin with segments having abrupt onsets, such as stops and affricates and it is uncommon for words to begin with gradual onset segments such as vowels [4].

Temporal Modulation Characteristics

- a. Micro-modulation patterns with periodicity between 3 and 12 ms

Most of the speech signal is produced while the vocal folds are vibrating. The acoustic result is that the speech waveform is modulated at a quasi-periodic rate ranging between 3 (330 Hz) and 12 ms (85 Hz). The lower limit is characteristic of the high range of female voices while the upper limit is typical of the low end of the male range.

- b. Macro-modulation patterns on the time scale of 50 to 250 ms

Riding on top of this micro-modulation is a longer periodicity associated with segments and syllables. The durational properties of syllables is a property of the permissible sound sequences in the language [5].

It is likely that all spoken languages are characterized by these properties, despite the differences in their phonetic inventories. It is these "universal" macro-properties of the speech signal that form the focus of the discussion below.

AUDITORY MECHANISMS

The auditory bases for the spectro-temporal properties described above are varied and complex, reflecting general constraints imposed by the mammalian acoustic transduction system. However, these general mammalian properties have special consequences for the nature of speech as a consequence of the unique fashion in which sensory-motor and cognitive information is integrated in the human brain.

Spectral Energy Distribution

Human listeners are sensitive to frequencies between 0.05 and 18 kHz [6]. However, the truly sensitive portion of this range lies between 0.25 and 8 kHz [6], setting approximate limits to the bandwidth of the acoustic communication channel.

The precise shape of the human audibility curve is conditioned by several factors. The lower branch of the audibility curve reflects the impedance characteristics of the middle ear [7]. The coupling of the ear drum to the oval

window of the cochlea via the ossicular chain is much more resistive to low frequencies than to high.

The upper branch of the audibility range is determined by the mechanics of the inner ear, which in turn is accounted for by macro-evolutionary factors pertaining to the ability to localize sound. The upper audibility limit pertains to the frequency range over which interaural intensity cues are available for precise localization of sound in the azimuthal and vertical planes. Because of the relatively large diameter of the human head (25 cm), it is possible to extract reliable localization information based on differential intensity cues for frequencies as low as 4-6 kHz [8]. This is an important limit, because the upper boundary of audibility for mammals is conditioned largely by the availability of these cues. For a small headed mammal, such as a mouse, comparable cues are available only in the human ultrasonic range, well above 20 kHz. Small headed mammals tend to be sensitive to far higher frequencies than their larger-headed counterparts [9]. Humans and other large-headed mammals need not be sensitive to the high-frequency portion of the spectrum since they can exploit both interaural time and intensity cues at the lower frequencies. In view of the limited number of neural elements available for frequency coding, an increase in bandwidth sensitivity necessarily reduces the proportion of tonotopic real estate focused on the lower portion of the spectrum. The comparative evidence suggests that there is a preference for focusing as much of the computational power of the auditory pathway on the lower end of the spectrum as possible, and that sensitivity to the high-end of the frequency range is a drawback except for the rather necessary function of source localization and characterization. There is a further implication, as well, that there is something special about the low-frequency portion of the spectrum, as discussed below.

Thus, it is no mystery why speech sounds contain little energy above 10

kHz. The human ear is relatively insensitive to frequencies above this limit because of the low-frequency orientation of the inner ear. But what precisely accounts for this low-end spectral bias, and could this account for the concentration of speech energy in the speech signal below 2.5 kHz?

One of the common properties of all vertebrate hearing systems is the ability of auditory neurons to temporally encode low-frequency information. This encoding is performed through a process referred to as "phase-locking," in which the time of occurrence of a neural discharge is correlated with the pressure waveform driving the cell. The limits of this temporal coding in the auditory periphery are a result of the modulation of neurotransmitter released by inner hair cells and the uptake of these chemicals by auditory-nerve fibers [10]. Auditory-nerve fibers phase-lock most effectively to frequencies up to 800 kHz, progressively diminishing in their temporal encoding potential with increasing frequency. The upper limit of robust phase-locking is ca. 2.5 kHz [11], although a residue of phase-locking persists up to 4-5 kHz [11].

What is the significance of phase-locking for encoding speech-relevant information? It provides a means of encoding spectral information in a robust, fault-tolerant fashion that is important for signal transmission in uncertain and potentially adverse acoustic conditions. In order to understand how this occurs, it is helpful to first consider another means of encoding frequency information in the auditory system.

The discharge rate of a peripheral auditory neuron is roughly proportional to the energy driving the cell over a limited range of sound pressure levels. Because of the filtering action of the cochlea, it would be possible to encode the spectrum of a complex signal entirely in terms of the average discharge rate of spectrally tuned neurons across a tonotopically organized neuronal array if the dynamic range of these cells was sufficiently large. However, the range

over which most auditory neurons increase their firing rate is only about 20-30 dB [12], far too small to effectively encode spectrally dynamic features. As a consequence, this "rate-place" representation of the acoustic spectrum becomes progressively blurred at higher sound pressure levels, despite robust perceptual decoding [13].

A significant problem with the rate-place code is its vulnerability to acoustic interference. Because the magnitude parameter is based simply on average rate it is impossible to distinguish neural activity evoked by a target signal of interest and extraneous background noise. There is no effective means of segregating the two sound sources on the basis of neural "labeling." A rate-place representation is particularly vulnerable to acoustic interference since its only measure of spectral identity is the location of activity in a topographically organized plane. Any extraneous source which intrudes into the place of the target signal could potentially disrupt its representation.

In contrast, phase-locked responses do provide a considerable degree of noise-robustness since the neural response is effectively labeled by the driving stimulus. A 500-Hz signal evokes a temporal distribution of nerve impulses rather distinct from one centered at 900 Hz, or even 550 Hz. This temporal information allows the auditory system to successfully segregate signals derived from disparate sources. In this fashion the temporal code provides a measure of protection against background sounds.

Phase-locking also provides a means to reduce the background noise as a consequence of its limited dynamic range. Frequencies more than 10-15 dB below the spectral peak driving the cell are rendered effectively invisible, since they do not affect the temporal discharge pattern [14]. This property effectively acts as both an automatic gain control [15] that suppresses background noise and enhances local peaks in the spectrum. This peak enhancement is combined with a broadening of the cochlear filters at

moderate to high sound pressure levels to spread the labeled low-frequency information over a wide tonotopic range of neurons. As a consequence, there are many neural channels carrying similar temporal information, and this redundancy of central nervous system input provides for a large measure of reliability in encoding such phase-locked information. Noise tends to be concentrated in particular frequency bands, and therefore its potentially disruptive effect minimized by virtue of the important information distributed over a large number of channels.

Because robust phase-locking only occurs for frequencies below 2.5 kHz, there is an advantage in using the lower portion of the spectrum for encoding information that needs to be transmitted over noisy acoustic channels. Thus, there is a real incentive from an information reliability perspective to pack as much information in the lower spectral bands as possible. Disruption of the representation can be detected and patched up because it is possible to associate similarly labeled activity. Furthermore, the limited dynamic range of phase-locking makes it more difficult for noise to disrupt the encoding of low-frequencies. The signal to noise ratio must be exceedingly low before auditory neurons lose the ability to phase-lock to the foreground signal.

Enhancement of spectral peaks in the neural activity pattern has the effect of center clipping in the frequency domain providing considerable incentive to concentrate communicationally relevant information in the low-frequency spectral peaks, particularly at high sound pressure levels.

Moderate to High Sound Pressure Levels

The energy in the speech signal is distributed as follows.

The voiced speech sounds, particularly the vowels and glides typically possess a considerable amount of energy, in the range of 60-75 dB SPL at the receiver's ear [4]. And most of these voiced segments concentrate their energy below 2.5 kHz [5]. The neural

signature cast by these segments spread their spectral shadow over much of the auditory nerve, recruiting high-frequency neurons to their temporal discharge, thereby rendering the information encoded relatively impervious to acoustic interference [16].

Segments with energy concentrations in the mid- and high-frequency (3-8 kHz) portions of the spectrum are typically of much lower amplitude (30-50 dB). This is probably a consequence of two factors. This is the most sensitive portion of the human ear's audible range. However, the gain in sensitivity is modest (10 dB) relative to frequencies between 0.5 and 2 kHz. A second factor is perhaps of more significance. This is the portion of the spectrum above the limits of neural phase-locking. Because of the limited dynamic range of auditory neurons the spectral contours associated with these mid-frequencies are most clearly delineated in the rate-place representation at low sound pressure levels. For this reason, these segments are more intelligible at low-to moderate intensities than at higher SPLs.

One may therefore surmise that the amplitude of individual classes of speech segments depends at least partially on the robustness of the resulting auditory representation, and not just on purely articulatory considerations. Low sound pressure levels are required for optimum encoding of high frequency spectra as they are dependent on rate-place auditory code which has a restricted dynamic range. Low-frequency spectra, on the other hand, are most robustly encoded at moderate-to-high sound pressure levels as a result of the spread of phase-locked excitation at these intensities.

Rapid changes in the Spectral Distribution of Energy over Time

a. Significant changes in spectral energy maxima over 100 ms intervals

One of the most salient acoustic properties of speech is its constantly changing character. Stop consonants come on abruptly. The formant patterns of liquids, glides and even vowels are constantly shifting. In continuous speech,

spoken at a normal rate, it is difficult to locate steady-state segments. This dynamic property of the speech signal is commonly interpreted as a reflection of a vocal tract pattern constantly in motion. And yet it is possible to produce quasi-steady-state speech-like segments in a continuous fashion, such as done in many forms of vocal music. But we don't typically communicate in chant.

So what factors account for the pace of spectral movement in the speech signal?

One consideration concerns the impact steady-state spectra have on the activity level of auditory neurons. A salient property of auditory neurons is adaptation. Turn on a signal and an auditory nerve fiber will fire at a very high rate for 5 to 15 ms [17], diminishing its activity level steadily over the next 100 ms or so. At higher levels of the auditory pathway many cells will fire only at signal onset. But this adaptation is highly frequency selective [18]. Change the spectrum of the signal and formerly quiescent neurons will turn on. In this fashion dynamic spectral properties could be used to maintain a high neural activity level. This preference for spectral dynamics is enhanced at the higher stations of the auditory pathway, where many cells respond only to spectrally changing stimuli, not to steady state signals.

A second factor pertains to the modulation transfer function of auditory cortical neurons, as discussed below. And a third factor pertains to the rate of information transmission through the auditory pathway into the central cortical areas, also discussed below.

b. Prevalence of abrupt onsets (e. g. stop consonants, clicks, affricates)

Abrupt onsets act in a manner similar to spectrally dynamic signals in that they tend to evoke a high rate of neural activity. This is a consequence of the fact that many cells in the central auditory pathway receive inputs from a wide tonotopically organized array of input neurons [19]. These cells act like "coincidence" detectors, responding to

stimulation only when a sufficient proportion of their inputs fire within a very small interval of time, typically 250 μ s or less. Among the more effective stimuli for simultaneous activation of a large neural array are transients associated with stop and affricate consonants, and as such are relatively more reliably encoded under adverse conditions.

Temporal Modulation Characteristics

- a. Micro-modulation patterns with periodicity between 3 and 12 ms.

A significant proportion of speech, perhaps as much as 80 percent [3], is produced while the vocal folds are vibrating. And yet this predominance of voicing is not necessary for sustained vocal production, as evidenced by whispered speech.

The fundamental frequency of adult human speech ranges from 75 Hz for a low, adult male to 330 Hz for a high female voice [5]. Although this range reflects to a certain degree the length and mass of the vocal folds, it is possible for the human voice to go well above this range, as attested by operatic performance. What is there about the normal f_0 range that makes it special with respect to the auditory coding of speech?

If we look at the ability of auditory neurons to encode the waveform periodicity of spectrally complex signals such as speech, phase-locking to this temporal parameter of the speech signal is most robust among central auditory neurons in the range 75-300 Hz [20] and is the region of the most acute modulation discrimination [21].

The significance of encoding waveform modulation becomes apparent when we consider how the auditory system would track a sound source through time without some equivalent form of cohesive force to bind disparate spectral elements together into a single sound source. Because speech and other communication signals are typically broadband, the system needs to know that the activity in the low-frequency channels is related to that evoked in the higher channels. Common periodicity

provides a cohesive cue that enables the system to attribute disparate neural activity to the same source.

Periodicity tracking in the range of the human voice is a particularly effective means of robust encoding of information [22]. At the level of the auditory nerve, fibers are capable of firing at sustained rates up to 250 spikes per second [23]. And at the level of the cochlear nucleus some cells can fire at rates up to 1000 spikes per second [24]. This phase-locked response to the modulation cycle enables each glottal cycle of the speech waveform to be marked with a burst of excitation that enables the system to track the signal across frequency and time. [25].

- b. Macro-modulation patterns on the time scale of 50 to 250 ms

In addition to the modulation of the speech waveform imposed by vibration of the glottis, is a slower amplitude modulation that is correlated with the passage of individual segments and syllables. A normal speaking rate (at least for English-speaking Americans) is approximately 4 syllables per second [4]. And, in English, there are approximately 2.5 segments per syllable [3]. Thus, the average length of a phonetic segment is ca. 100 ms and that of a syllable, 250 ms.

At the higher stations of the auditory pathway, principally the auditory cortex, neurons generally respond at rates between 5 - 20 Hz (50 - 100 ms) [26]. Each cell in this region acts as an integrative center, its response reflecting the activity of hundreds, if not thousands of cells at more peripheral stations. It appears likely that the syllable and segment rate of spoken discourse is at least partially conditioned by the firing rates of these cortical neurons. These cells can phase-lock to the slow modulations of energy within their response areas and thereby provide an accurate representation of both syllabic and gross spectral contours. The gross amplitude modulation cues can be used to temporally bind neural activity driven by different portions of the spectrum, but having common onsets and offsets.

CORTICO-THALAMIC MECHANISMS

The brain utilizes specific strategies to integrate auditory information into linguistically meaningful units. It is rather unlikely that speech is processed phoneme by phoneme like so many "beads on a string." Rather the acoustic components of speech appear to segregate into syllable or mora-length units, which in turn are integrated into words and higher level semantic units.

What are the neural bases for this syllable timing properties of speech?

Many properties of auditory function appear to be governed by a 200-ms time constant, including temporal masking, intensity integration and loudness summation [6]. It is of interest that a similar time constant figures prominently in visual and motor function as well [27].

These observations suggest the existence of a quantal unit common to the sensory and motor systems, a unit of time over which sensory information is analyzed and correlated with the relevant motor systems, possibly through the reticular nuclear complex of the thalamus and the neo-dentate nucleus of the cerebellum.

Thus, the syllable may serve as the temporal unit for integration of auditory information into higher-level linguistic units.

During speech production the motor system controlling the vocal apparatus is almost surely in close communication with the output of the auditory system. The syllable may thus serve as the temporal unit for which the auditory and articulatory components of speech are synchronized, and also serve as well as the basic unit for higher level integration into semantic units.

Information encoded in syllable packets places a temporal constraint on linguistic information. It establishes this time frame as one in which a minimum amount of information needs to be fit for higher level integration.

These observations suggest the existence of a quantal unit common to the sensory and motor systems, a unit of time over which sensory information is analyzed and correlated with the relevant

motor systems, probably through the reticular nuclear complex of the thalamus.

SUMMARY AND CONCLUSIONS

The human vocal apparatus is likely to have evolved under conditions optimized to produce communication signals possessing properties that exploit the auditory system's ability to encode information in a robust, fault-tolerant fashion.

The speech spectrum is biased towards the low-frequencies which are particularly resistant to disruption from background noise. The sound pressure level of most speech is sufficiently high as to insure that low-frequency spectral information is spread across a wide array of auditory frequency channels. Glottal periodicity insures that the system is able to track speech in noisy, acoustically adverse conditions, and syllable length modulation helps the brain bind together disparate spectral entities into meaningful units.

Within this framework, the importance of the auditory system for speech is that it preconditions the neural representation for maximum reliability and rate of information transmission. It does this by creating a sparse representation of the signal, consisting mainly of changes in spectral peaks and temporal parameters. The brain therefore only needs to keep track of novel features in the waveform, confident that only these encode important information.

Is this correlation between auditory properties and the speech waveform sufficient to fully account for the acoustic properties of human language? Probably not. Although the auditory system necessarily provides the sort of robust, efficient form on information representation required for higher level linguistics integration, it fails to fully specify why speech occurs in syllable and word level units.

Other brain centers, such as the thalamus and cortical association areas are undoubtedly involved in the transformation of this acoustic information into a complex symbolic system.

REFERENCES

- [1] Pickett, J. (1980), *The sounds of speech communication*, Baltimore: University Park Press.
- [2] O'Shaughnessy, D. O. (1987), *Speech communication: human and machine*, Reading, MA: Addison-Wesley.
- [3] Miller, G. (1951), *Language and communication*, New York: McGraw-Hill.
- [4] Fletcher, H. (1953), *Speech and hearing in communication*, New York: van Nostrand.
- [5] Lehiste, I. (1970), *Suprasegmentals*, Cambridge, MA: MIT Press.
- [6] Moore, B. C. J. (1989), *Introduction to the psychology of hearing* (3rd ed.), London: Academic Press.
- [7] Dallos, P. (1973), *The auditory periphery: biophysics and physiology*. New York: Academic Press.
- [8] Erulkar, S. D. (1972), "Comparative aspects of spatial localization of sound", *Physiological Reviews*, vol. 52, pp. 237-360.
- [9] Masterton R. B.. (1974), "Adaptation for sound localization in the ear and brainstem of mammals", *Federation Proceedings*, vol. 33, pp. 1904-1910.
- [10] Palmer, A. R and Russell I. J. (1986), "Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells", *Hearing Research*, vol. 24, pp. 1-15.
- [11] Johnson, D. (1980), "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones", *J. Acoust. Soc. Am.*, vol. 68, pp. 1115-1122.
- [12] Palmer A. R. and Evans E. F. (1980), Cochlear fibre rate-intensity functions: no evidence for basilar membrane nonlinearities, *Hearing Research*, vol. 2, pp. 319-326.
- [13] Sachs M. B. and Young E. D. (1979), "Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate", *J. Acoust. Soc. Am.*, vol. 66, pp. 470-479.
- [14] Greenberg S, Geisler C. D and Deng, L. (1986), Frequency selectivity of single cochlear-nerve fibers based on the temporal response pattern to two-tone signals, *J. Acoust. Soc. Am.*, vol. 79, pp. 1010-1019.
- [15] Geisler C. D and Greenberg S. (1986), "A two-stage nonlinear cochlear model possesses automatic gain control", *J. Acoust. Soc. Am.*, vol. 80, pp. 1359-1363.
- [16] Young E. D. and Sachs, M. B. (1979), "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers", *J. Acoust. Soc. Am.*, vol. 66, pp. 1381-1403.
- [17] Smith R. L. (1977), "Short-term adaptation in single auditory nerve fibers: some poststimulatory effects", *J. Neurophys.*, vol. 40, pp. 1098-1111.
- [18] Harris D. M. and Dallos P., (1977), "Forward masking of auditory nerve fiber responses", *J. Neurophys.*, vol. 42, pp. 1083-1107.
- [19] Rhode, W, S, Oertel D. and Smith P. H. (1983) Physiological response properties of cells labeled intracellularly with horseradish peroxidase in cat ventral cochlear nucleus. *J. Comp. Neurol.*, vol. 213, pp. 448-463.
- [20] Langner, G and Schreiner C. E. (1988) Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *J. Neurophys.*, vol. 60, pp. 1799-1822.
- [21] Bacon S. P. and Viemeister N. F. (1985) "Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners", *Audiology*, vol. 24, pp. 117-134.
- [22] Bregman, A. (1990) *Auditory scene analysis*. Cambridge: MIT Press.
- [23] Kiang, N. Y.-S. (1965) *Discharge patterns of single fibers in the cat's auditory nerve*. Cambridge, MA: MIT Press.
- [24] Rhode, W. S. and Smith, P. H. (1986) "Encoding timing and intensity in the ventral cochlear nucleus of the cat", *J. Neurophys.*, vol. 56, pp. 261-286.
- [25] Holton, T. (1995) "A computational approach to recognition of speech features using models of auditory signal processing", *Proc. Int. Cong. Phon. Sci.*
- [26] 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 Research*, vol. 21, pp. 227-41.
- [27] Kandel, E. R. and Schwartz, J. H. (1985) *Principles of neural science* (2nd ed.) New York: Elsevier.