EECS 225
Lecture 14
February 17, 1995

Auditory and Perceptual Constraints on the Spectro-temporal Properties of Speech

Steven Greenberg

14.1 Summary

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

The auditory parameters of interest are: (1) the range of frequency sensitivity, (2) the frequency resolution of the inner ear, (3) the limits of neural periodicity coding, (4) the time course of rapid adaptation and (5) the temporal limits of neural coincidence detection.

In the human these parameters account for the following acoustic properties of speech: (1) the absence of spectral energy above 10 kHz, (2) the concentration of spectral information below 2 kHz, (3) a preference for quasi-periodic waveforms whose fundamental frequencies range between 75 and 400 Hz and (4) the prevalence of phonetic segments with rapid initial onsets.

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 (the primordial cocktail party).

The higher cortical centers impose constraints of a different sort, namely those pertaining to the rate at which acoustic information is converted into semantic form (and vice versa). Such conversion involves readout from memory, a highly time-dependent process.

Although speech shares many gross spectral properties in common with non-human vocal communication systems (spectral resonances, quasi-periodicity, rapid onsets, prevalence of low-frequency spectral energy), the temporal parameters of language are very different. Whereas the vocalizations of non-human species are largely repetitive or occur in relative isolation, speech is characterized by the chaining of heterogeneous acoustic segments of relatively short duration (50-200 ms) into syllable units occurring at a rate of two to eight per second. In order for the speech stream to be reliably decoded, higher cortical centers must be capable of converting phonetic information into semantic form with speed and precision. This conversion process is likely to involve neural populations in the parietal and pre-frontal regions, both of which underwent significant expansion in the evolutionary transition from ape to hominid. Without such cortical expansion it is unlikely that acoustic signals could have been decoded at a rate typical of human speech.

Although the auditory component of the language-processing system is not unique to humans, its integrity is essential for higher-level cortical processing to succeed. Because of the speed at which the phonetic-to-semantic conversion occurs, the auditory representation of the acoustic input must be sufficiently robust as to insure reliable decoding on a single presentation. The key to understanding the origins of human language lies in the interface between these two systems and their articulation with the vocal apparatus.

14.2 Introduction

In this presentation I would like to depart from the traditional perspective and focus on a component of the system that is shared in common with our primate and mammalian relations, namely the auditory system. This focus is intended to provide some insight into why languages sound the way they do and explore the implications for the higher cortical centers responsible for linguistic processing. Although this perspective will not tell us precisely how language evolved, hopefully it can provide some insight into the form of behavioral and neural constraints that shaped the development of our species vocal communication system.

I'll also depart from tradition in another way. Traditionally, the acoustic structure of speech is considered to primarily reflect properties of the human vocal apparatus. The tongue, lips and jaw can move only so fast, the size and shape of the vocal tract govern the spectral configuration of speech, etc. Although there is little doubt that articulatory properties impose real constraints on the range of acoustic possibilities, it is unlikely that these factors, by themselves, can adequately account for the specific constellation of spectro-temporal properties of speech. There are sounds which the vocal apparatus is capable of producing, such as coughing, spitting, blabbering, etc. which do not occur in any language's sound inventory. And although speech can be whispered, it is rarely done. The vocal tract is capable of chaining long sequences of vowels or consonants but no language relies exclusively on a single type of speech segment, nor does speech contain long sequences of acoustically similar segments.

The central thesis of my presentation is that auditory system is crucial for understanding how information is packaged in the speech waveform, and that its properties were instrumental in guiding the evolution of our species' vocal communication system. However, auditory constraints, in and of themselves are inadequate for understanding the origins of language. For this we must look to the higher cortical association areas, particularly in the frontal lobes, where the neural residue of the acoustic information is integrated into higher-order linguistic units.

14.3 Spectro-temporal Properties of Speech

The major spectro-temporal properties of speech of interest to our discussion are:

1. *Spectral energy distribution*

A. Entirely below 10 kHz

There is no speech segment which contains significant energy greater than 10 kHz. And most speech segments in fact contain little energy above 4 kHz. It is well-known from studies of telephone speech that only spectral information below 3.4 kHz is required for nearly perfect intelligibility.

B. Concentration of energy below 3 kHz

The long term power spectrum of the speech signal is concentrated below 3 kHz, due principally to the temporal domination of vowels, glides and liquids in the speech stream.

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 10-30 dB lower for stops and fricatives.

3. *Rapid changes in the 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.

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.

To my knowledge, 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 we will be concerned with in this presentation.

14.4 The Auditory Framework

We now turn our attention to the auditory system, and specifically to how its properties can account for the spectro-temporal features of speech.

Most of the remainder of the presentation will focus on speech properties 1 through 6. The latter part of my presentation discusses the significance of the seventh property for integration into higher level units of meaning.

Overview

The auditory pathway is a complex system involving dozens of anatomical structures and millions of nerve cells. For the purposes of this presentation we focus on five anatomical regions: the middle ear, the cochlea, the auditory nerve, the cochlear nuclei and the auditory cortex.

We discuss the auditory basis of each of the acoustic attributes of speech enumerated above:

14.4.1 Spectral Energy Distribution

Humans are sensitive to frequencies between 0.05 and 18 kHz. However, the truly sensitive portion of this range lies between 0.5 and 4 kHz. 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. 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 are accounted for by macro-evolutionary factors pertaining to the ability of localize sound.

There is a very limited amount of "real estate" in the inner ear for transducing spectral information. Space devoted to one part of the frequency range necessarily effects the resolution in other regions of the spectrum. Because of the gradient properties of the cochlea, a strategic decision must be made as to how to apportion the real estate available across the frequency spectrum. For reasons that will become clear soon, small headed mammals, such as rodents, tend to allocate a greater proportion of there cochlear real estate to the high frequency portion of the spectrum, while large headed mammals, such as humans and elephants, focus on the low-end of the spectrum (certain predators such as the cat cover both ends about equally well).

The reason for this strategic distinction in spectral allocation of neural topography is evident from a consideration of the two primary types of physical cues available to the brain for locating a sound source.

Imagine that you a looking straight at a singing bird which is perched on a branch level with your eyes. Close your eyes and listen to the bird's song. Now open your eyes and listen. There is something about the sound tells us that the sound is coming from straight ahead. Now, turn your head to the right, away from the bird. Close your eyes again, and listen. Notice how the sound has changed from the first condition?

The cues that correlate with position of sound in the horizontal (azimuthal) plane are the disparity in sound pressure level and time of arrival to the two ears. Interaural time cues are more reliable and accurate than interaural intensity cues, but require a sufficiently large head for the time differences to be long enough to be reliably encoded in the auditory pathway. Virtually all large-headed mammals rely principally on interaural time cues for azimuthal localization. For large-headed mammals, there is a predictable relationship between head size and the frequency range exploitable for localization using interaural intensity cues. In order for large intensity disparities to occur the wavelength of the sound must be small relative to the obstruction, in this instance the head. Wavelength is inversely proportional to frequency. The wavelength of a 1 kHz signal is approximately 1 foot, while that of a 50 Hz signal is 20 feet. (1130 ft per second/ 1000 Hz ~ 1 foot). For an obstacle such as the head to cast an effective sound shadow its diameter needs to be at least four time greater than the signal's wavelength. For the human head, with an average diameter of 10" this means that only frequencies above 6 kHz will be attenuated sufficiently to provide reliable interaural intensity cues. For a small-headed species such as the mouse, the same relationship holds, but scaled up in frequency by a factor of 5-10.

Thus, 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. And the lower the upper limit of audibility, the greater proportion of cochlear real estate available for encoding low-frequency signals, the significance of which will be discussed shortly.

So it is no mystery that speech sounds do not contain energy above 10 kHz. The human ear is relatively insensitive to frequencies above this limit, which is a product of evolutionary pressures pertaining to sound localization and the frequency allocation of the inner ear.

But what accounts for the concentration of energy in the speech signal below 3 kHz?

We need to delve much deeper into the nature of frequency coding to understand this emphasis on the low-frequency portion of the spectrum. There are special physiological mechanisms that enable signals with low-frequency energy to be encoded in a robust, fault-tolerant fashion of particular importance for signal transmission in acoustically uncertain and potentially adverse conditions.

There are two basic means of encoding frequency information in the auditory system. At low sound pressure levels a narrow-band signal will evoke a higher amount of neural activity at a particular location in the cochlea and other tonotopically organized loci. Low frequency signals evoke a maximum level of activity at locations near the apex of the cochlea while high frequencies evoke a maximum near the base. For reasons that will be discussed later on, this "place" mode of frequency coding can not account for the full range of frequency coding capabilities that humans and other mammals are capable of.

The second means of frequency coding is based on the temporal distribution of neural activity. For low-frequency signals there is a systematic relationship between temporal pattern of neural discharge and the stimulating waveform. Individual discharges can be traced

to specific points in time of the waveform. We called these responses "phase-locked" since they are highly correlated with a particular phase (time) portion of the signal.

Neurons in the auditory periphery can phase-lock to frequencies up to about 4 kHz, though it is unusual for strong neural synchrony to be evident above 2 kHz.

This temporal mode of frequency coding occurs in tandem with the place mode, and usually the two operate in concert.

A drawback of the place code is that it 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.

The temporal code provides some protection against background sounds since the neural response is "labeled" in terms of the temporal intervals between discharges, that acts as a signature of the signal driving the cell. Disruption of the representation can be detected and patched up because it is possible to associate similarly labeled activity. In addition, 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.

Another consequence of phase-locking is to suppress the portions of the spectrum outside of the peaks. This is also due to its limited dynamic range. Signals which are more than 10 dB from the peak are not at all represented in the neural response. This has the effect of center clipping in the frequency domain, which enhances the representation of the spectral peaks relative to the troughs - precisely what appears to occur for speech. This is achieved by spreading the low-frequency peak information across the topographically organized array of neurons at multiple levels of the auditory pathway.

To understand how this occurs let's examine the response functions of auditory-nerve fibers. These response areas and tuning curves measure a cell's ability of a cell to modulate its excitability on the basis of a signal's spectral composition. The classic picture of a cell's tuning function is shown in the accompanying illustration.

It is sharply tuned to a narrow range of frequencies and is unresponsive outside this range. This picture is accurate only for cells most sensitive to frequencies below 3 kHz. Above this frequency range, the area of the cell's response area broadens enormously at moderate to high sound pressure levels, as shown in the illustration. The significance of this is that cells which respond only to high frequency inputs near their threshold, can respond to low-frequency signals at higher SPLs.

Without phase-locking there would be no distinctive signature low-frequency sounds could impart on neurons in high-frequency channels. However, because of the synchronization to the driving stimulus the system is able to distinguish the frequency of the driving stimulus. Thus the place principle of frequency coding is supplanted to a large degree at high SPLs by a temporal coding scheme that enables more than a single frequency channel to encode low-frequency spectral information. This provides a powerful means of distributing important information across frequency channels in the cochlea and beyond that could increase the reliability of acoustic information transfer dramatically.

However, to take advantage of these singular properties the information must be encodable in such a form; in other words must be concentrated in the frequency bandwidth below 3 kHz. There is tremendous pressure from an information reliability perspective to pack as much information in the lower spectral bands as possible. This, I suggest is the principle reason why so much of the speech signal's energy is below 3 kHz.

14.4.2 Sound pressure levels between 30 (stops) and 70 dB (vowels) for naturally spoken speech

The energy in the speech signal is distributed in the following manner:

The voiced speech sounds, particularly the vowels and glides have SPLs in the range of 60-75 dB. And these signals have energy concentrated primarily below 3 kHz. It is these segments which would cast their spectral shadow over much of the cochlear partition, recruiting high-frequency neurons to their temporal discharge, and thus be less vulnerable to acoustic interference.

Segments with their energy concentrations in the mid- and high-frequency portions of the spectrum are of much lower amplitude. This is probably a consequence of two factors. First, this is the most sensitive portion of the spectrum, due to the resonances of the external ear. Second, this is the portion of the spectrum above the limits of phase-locking. Because of the limited dynamic range of most auditory neurons the spectral peaks in this region of the spectrum can not be encoded in terms of a place representation for sound pressure levels above 50-60 dB. For this reason, these segments are more intelligible at low-to moderate intensities than at higher SPLs.

Therefore, it would appear that the primary factor governing the amplitude of individual classes of speech segments is the robustness of the auditory representation, not articulatory considerations. For it is certainly possible to articulate vowels at much lower intensities and fricatives and stops at much higher amplitudes than is done in normal continuous discourse (One has only to think of Barbara Woodhouse's verbal training regimen for wayward canines to be convinced of this).

14.4.3 Rapid changes in the spectral distribution of energy over time

A. Significant changes in spectral energy maxima over 100 ms intervals
B. Prevalence of abrupt onsets (e. g. stop consonants, clicks, affricates)

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 any steady-state segment. This dynamic property of the speech signal is traditionally interpreted as a reflection of the vocal tract, which is constantly in motion. However, it is possible for the vocal tract to produce quasi-steady-state signals in a continuous fashion, such as done in many forms of vocal music.

So why don't we normally communicate in chant?

One possible factor pertains to the rate of information transmission, and we will return to this issue towards the end of our presentation.

Another potential factor involves 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, diminishing its activity level steadily over the next 100 ms or so. At the higher levels of the auditory pathway many cells will fire only at signal onset. But this adaptation is highly frequency selective. Change the spectrum of the signal and formerly quiescent neurons will turn on. In other words, spectral dynamics is a way to maintain a high activity level during the course of the signal by continuously recruiting new cells for encoding the signal. At the higher stations of the auditory pathway this preference for changing spectra is reflected in the fact that many cells respond only to spectrally changing stimuli, not to steady states.

Many cells in regions such as the auditory cortex receive inputs from a wide spectral range of cells. These neurons act like "coincidence" detectors, in that they respond to stimulation only when a sufficient proportion of their inputs fire relatively simultaneously. The most effective stimuli for turning on a large array of cells at the same time are transients such as occur at the onsets of stop and affricate consonants.

Thus, a major consequence of speech's dynamic character is to increase the overall activity level of the auditory system. This increase in activity level almost surely results in increased reliability and robustness in the face of acoustically adverse conditions.

14.4.4 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, is produced while the vocal folds are vibrating. 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 300 Hz for a high female voice. Although this range reflects to a certain degree the size and mass of the larynx, it is possible for the human voice to go well above this range, as attested by operatic performance. So again, we ask what is there about the normal f_o 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, we find phase-locking to this temporal parameter of the speech signal is most robust in the range 75-300 Hz. And this is also reflected in behavioral data pertaining to the accuracy of discriminating small changes in modulation frequency.

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 the cohesive cue that enables the system to allocate separate neural activity to the same source.

Periodicity tracking in the range of the human voiced is also characterized by a particularly effective means of robust encoding of information. At the level of the auditory nerve, fibers are capable of firing at sustained rates up to 250 spikes per second. And at the level of the cochlear nucleus some cells can fire up to 600 spikes per second. This means that is possible to mark each glottal cycle of the speech waveform with a burst of excitation that enables the system to track the signal across frequency and time.

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.

American English is normally spoken at a rate of approximately 4 syllables per second. And there are on average 2.5 segments per syllable. This means that the average length of a phonetic segment is 100 ms and that of a syllable 250 ms. Of course, this is merely an average figure. Speaking rates can be slower and or much faster, but the interesting thing is that once the speaking rate exceeds a certain threshold its intelligibility plummets.

The important point is that the auditory system can utilize gross amplitude modulation cues to bind in time segments with different spectral features in an effort to parse the speech waveform for meaningful units.

At the higher stations of the auditory pathway, principally the auditory cortex, cells generally respond at rates between 5 - 20 Hz (50 - 100 ms). Each neuron 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 is tempting to speculate that the normative speaking rate is

conditioned, at least in part, by the properties of integrative neurons in the higher stations of the auditory pathway.

14.5 Recapitulation

In the first part of this presentation we have shown how the evolution of the human vocal apparatus is likely to have been shaped to produce communication signals with properties that exploit the auditory system's ability to encode information in a robust, fault-tolerant manner.

The spectrum of the speech is biased towards to 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? I think 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 and related structures are undoubtedly involved in the transformation of this acoustic information into a complex symbolic system. I will focus on these extra-auditory mechanisms for the remainder of my presentation.

14.6 The Cortical-Temporal Framework

The brain requires certain strategies to integrate spectro-temporal patterns information from the auditory pathway 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.

It is this temporal integration into higher level units that in my opinion most distinguishes speech from other vocal communication systems. This is not to imply that other communication systems utterly lack this form of higher level integration, but rather that the complexity and depth of the integration is likely to be significantly greater in our species.

But let us step back from language briefly to consider the possible origins of the auditory cortical preference for stimulus rates of 4 to 5 per second.

Many sensory systems function as energy integrators. This means that the magnitude of response is proportional to signal duration - up to a certain length, and then saturates. In the auditory system this integration period is around 200 ms, and is similar to the length of energy integration in the visual system.

In the auditory system certain forms of perceptual sequences appear to be governed by this integration time constant. For example, if one switches between two signals, whose frequencies are relatively close, then if the durations are less than 250 ms, one hears under the appropriate

conditions a continuous tone and a pulsing tone, indicating that energy from the two signals has been integrated into a single object. Analogous phenomena occur in the visual modality.

The motor system also appears to have a finite time constant for activation. It has been shown that a motor act can not be aborted or modified within 250 ms of its initiation, as if the period preceding it is part of the motor act itself.

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.

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 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.

Another, not entirely unrelated constraint on temporal processing is short-term memory. In the evolution of the memory system it is likely that temporal integration times were matched to the pace at which the sensory and motor systems could respond.

Information that occurs at too slow a pace is likely to be easily integrated into higher level units due to a lack of temporal coherence. For example, if speech is spoken at a very slow rate then words get broken into a series of syllables that lose their linguistic identity.

Thus, the temporal dimension of speech is likely to reflect in some measure the amount of information conveyed. Speech is spoken at a rate calculated to insure its accurate interpretation. Complex messages are spoken slowly, while more predictable utterances are usually produced at a faster rate.

In my view it is the brain mechanisms responsible for converting the auditory stream into a sequence of meaningful units that marks the transition from proto-human to human communication. While other mammals surely possess the ability to convert sound into some form of referential meaning, the complexity and diversity of the intended message is immeasurably greater.

Quantity of information storage and access speed are probably key parameters governing the evolution of the language portions of the brain. Large storage systems must be fast, otherwise the average access speed will be unacceptably slow. So an important development in the evolution of language must have been the ability to retrieve information rapidly.