Temporal Properties of Spoken Language

Steven Greenberg

The Speech Institute
Santa Venetia, CA 94903, USA
steveng@cogsci.berkeley.edu

Abstract

Time is frequently conceptualized as a dimension apart from others, binding disparate processes through its unidirectional flow. In this sense time is an abstraction providing a convenient perspective with which to analyze complex phenomena within a unified framework. With respect to speech, this framework pertains to information unfolding over time. Within communication it is essential for processes associated with a message's encoding and decoding to be synchronized in time. Because information lies at the foundation of speech communication, and because information is inextricably bound with time, the temporal dimension lies at the heart of spoken language. This close affinity between time and information affords keen insight into the very nature of speech, addressing such fundamental questions as: (1) why is speech spoken at specific rates, and what accounts for the variation in timing observed in daily conversation, (2) why is each organizational tier of spoken language associated with a distinctive span of time, and (3) what is the specific relation between time and information contained in the spoken message? Speech dynamics affords a fruitful framework with which to examine the relation between information and spoken language from the perspectives of perception, production, neurology and technology.

1. Introduction

Speech is inherently dynamic, reflecting the motion of the tongue and other articulators during the course of vocal production. Such articulatory dynamics are reflected in rapid spectral changes, often called formant transitions, characteristic of the acoustic signal. Although such dynamic properties have long been of interest to speech scientists, their fundamental importance for spoken language has only recently received broad recognition.

Traditionally, speech dynamics have been examined principally from a biomechanical perspective. Given the structural constraints imposed through phylogenetic descent, speech production has generally been viewed as nature's way of solving an exceedingly complicated problem with limited biomechanical means. The jaw, tongue, lips and other articulators can move only so fast, their rates of motion limited by their anatomical and physiological characteristics. Such properties reflect an evolutionary process long antedating the origins of human vocal communication. From this purely articulatory perspective, speech's spectro-temporal properties are primarily the consequence of biomechanical constraints imposed through the course of human (and mammalian) evolution.

If the fine details of spoken language are indeed governed by vocal production, how does the brain decode the speech signal given the acoustic nature of the input to the auditory system? One prominent model, known as "Motor Theory," posits that the brain back-computes the articulatory gestures directly from the acoustic signal [15]. In essence, this framework likens the auditory system to a delivery service that transmits packages containing articulatory gestures decoded at some higher level of the brain. The process of perceiving (and ultimately understanding) speech thus reduces to associating articulatory gestures with sounds and words.

A central issue for Motor Theory and other models of speech perception is the "invariance problem." Words maintain an essential identity independent of the speaker and the environment. The same word, spoken by the same individual, differs from one instance to the next as a consequence of variation in pronunciation and the acoustic environment. The word's spectro-temporal properties vary accordingly, and yet each lexical instance is ultimately interpreted as the same as other instances of that word. How does the brain learn to "ignore" such acoustic variation and focus on the essence of the message?

2. The Medium is the Message

An important issue in contemporary speech research concerns the manner in which information is packaged in the signal. As early as 1939, it was recognized that the meaningful component of speech ("intelligibility") is associated with slow variation in acoustic energy, spanning intervals between 40 and 400 ms [6]. This insight enabled Dudley to develop the first truly intelligible synthesizer, known as the VOCODER. Dudley was careful to note that intelligibility required not only slow variation in energy, but its *differential distribution* across the frequency spectrum [6]. In present-day terminology, we would characterize this insight as a distinction between the "modulator" and "carrier" components of the signal.

Dudley's bold perspective was initially taken most seriously by Chistovich and colleagues in Leningrad, beginning in the 1960's. This group was perhaps the

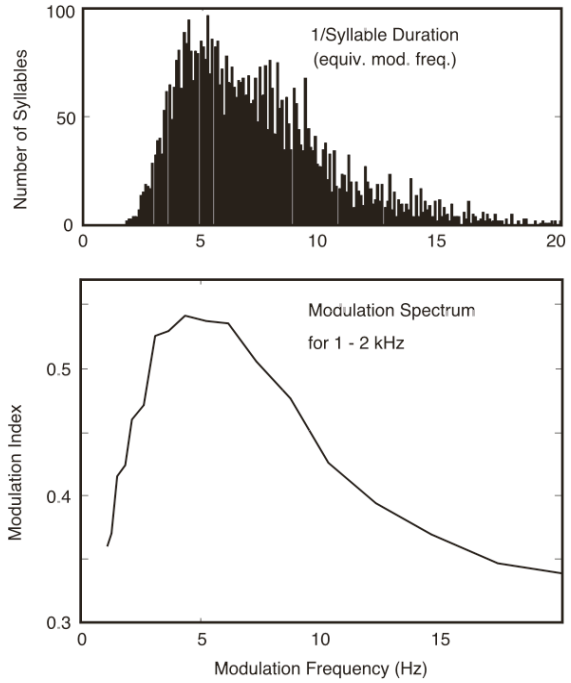

Figure 1: The relation between the distribution of syllable duration (transformed into modulation frequency) and the modulation spectrum computed for the octave region between 1 and 2 kHz for 15 minutes of spontaneous Japanese material. Adapted from [1]

first to recognize the intimate relation between production and perception that is mediated by linguistic units longer than the phone (e.g., the syllable), emphasizing the highly non-linear, dynamic nature of speech [14].

3. The Modulation Spectrum

The next major advance in speech dynamics was introduced by Houtgast and Steeneken in the 1970's. They computed the "modulation spectrum" as a means of predicting intelligibility in various acoustic environments, noting that rooms in which speech is easily understood possess a distinctive profile with respect to energy fluctuations in the acoustic signal [13]. The modulation spectrum's peak in highly intelligible environments was found to be ca. 4-5 Hz, with a broad distribution of energy between 2 and 10 Hz (see Figure 1, lower panel). They correctly noted that the modulation spectrum's peak conformed roughly to the average duration of syllables (see Figure 1, upper panel) and speculated that intelligibility depended on acoustic boundaries between adjacent syllables being well preserved. Consistent with this hypothesis was their observation that speech is difficult to understand under precisely those conditions where the magnitude of the low-frequency modulation spectrum is severely attenuated (as occurs at extremely low signal-to-noise ratios) or when its peak shifts below 2 Hz (as occurs in highly reverberant environments). In the mid-1990's Drullman tested such

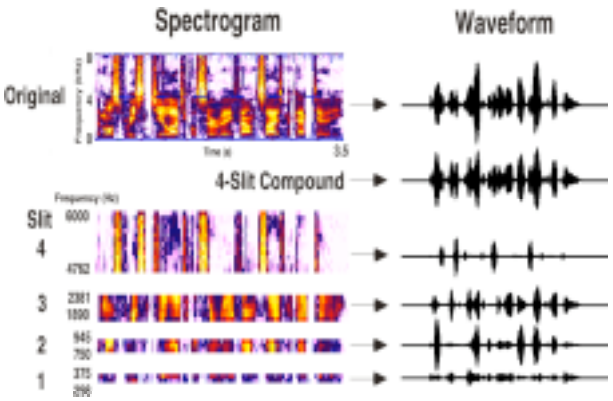

Figure 2: Spectrographic and time-domain representations of a representative sentence ("The most recent geological survey found seismic activity") used in an intelligibility study [11]. The frequency axis of the spectrographic display of the channels has been non-linearly compressed for illustrative tractability. Note the quasi-orthogonal temporal registration of the waveform modulation pattern across frequency channels. From [11].

assumptions in a series of elegant perceptual studies [4][5], demonstrating that intelligibility does indeed depend on the integrity of the modulation spectrum below 8 Hz.

4. Modulation's Differential Distribution Across the Acoustic Frequency Spectrum

In Dudley's original formulation of the VOCODER the modulator and carrier shared equal billing. Both were viewed as required for intelligibility. Dudley estimated that the *acoustic* spectrum needed to be partitioned into approximately 10 distinctive channels to produce high-quality speech [6]. In the mid-1990's Shannon and colleagues developed a variant of the VOCODER that cast considerable doubt upon this parity between carrier and modulator [16]. In their study only four separate channels were required to produce intelligible speech. Moreover, the carrier used was not harmonically structured, as would be found in voiced speech, but Gaussian noise, akin to the glottal source associated with whispering [16]. Shannon's result implied that the modulator is far more important than the carrier, and that the primary function of the acoustic spectrum is to serve as a medium with which to differentiate the dynamic characteristics of the modulator across the frequency (i.e., tonotopic) plane.

More recently, others have shown that intelligibility does not directly depend on the fine details of the frequency spectrum and can largely be dispensed with, as long as certain essential modulation properties associated with the original signal are preserved [2][3][9], and their temporal (i.e., phase) relation across the acoustic frequency plane maintained [10][11][17] (see Figure 2). Intelligibility appears to depend on *both* the magnitude and phase components of the modulation spectrum, as discussed in Section 5.

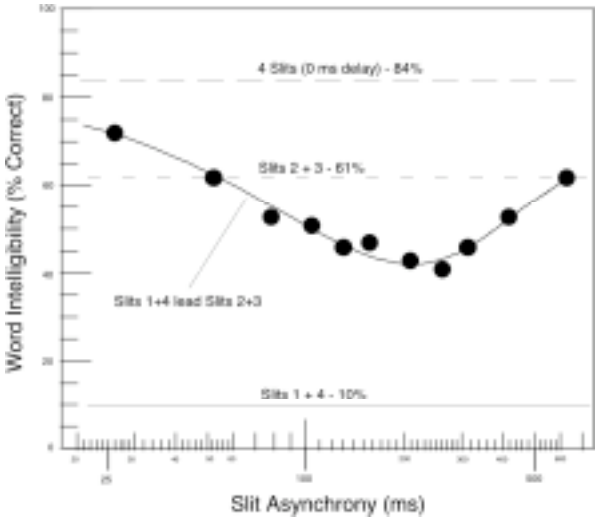

Figure 3: Intelligibility of sparse spectral sentences containing four narrow-band (1/3 octave) channels as a function of slit asynchrony. Note that intelligibility goes below baseline (slits 2+3) when the slit asynchrony exceeds 50 ms). 27 subjects. Adapted from [17]

5. The Importance of Modulation Phase

The importance of modulation phase across the tonotopically organized frequency spectrum can be demonstrated through a separate experiment, in which two narrow-band (1/3 octave) spectral slits in the central region of the acoustic spectrum (0.9 – 2.1 kHz) are systematically delayed in time relative to a very-low-frequency (335 Hz) and a very-high-frequency (5400 Hz) slit. When entirely synchronized, the four slits yield 84% intelligibility. As the two central slits lag their lateral counterparts by increasing intervals intelligibility diminishes progressively, reaching a trough between 250 and 300 ms of asynchrony (Figure 3). Longer lag times result in a slight rebound in intellibility (Figure 3).

The shape of the intelligibility curve approximates the statistical distribution of syllable duration in English (Figure 4, bottom panel), and is consistent with the hypothesis that very-low-frequency modulations in the acoustic signal reflects the syllabic structure of speech. Such syllable structure is likely to be similar across languages, as illustrated for spontaneous English and Japanese material in Figure 4 [1]. However, there is more to decoding speech than merely processing the low-frequency modulation spectrum of the acoustic signal.

6. Audio-Visual Integration of Speech

In the real world speech visual cues are often used as an important supplement to the acoustic signal, particularly in noisy backgrounds and among the hearing impaired [8]. The time constants pertaining to audio-visual integration of spoken language are shown in Figure 5. The structure of this perceptual experiment is similar to that illustrated in Figure 3. However, in place of the

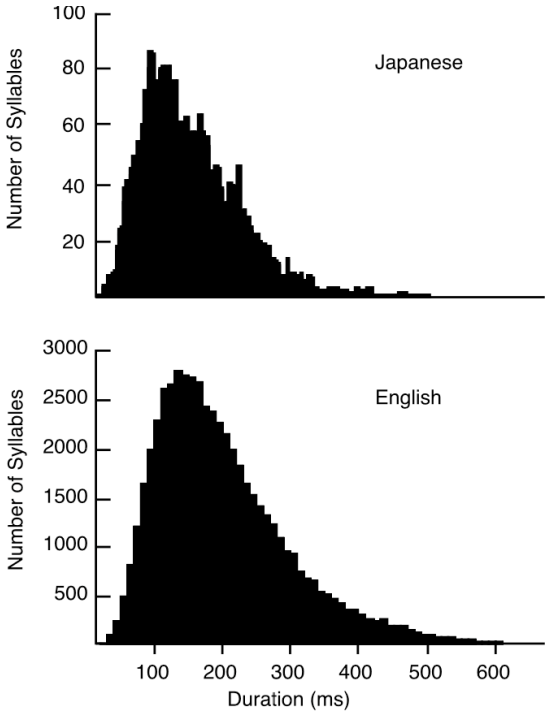

Figure 4: Statistical distribution of syllable duration for spontaneous material in Japanese and American English. Adapted from [1].

two acoustically central slits used in the earlier study, a video of the talker speaking each sentence is shown in tandem with presentation of the acoustically lateral (i.e., the 300-Hz and 5400-Hz) slits [7].

In the absence of non-visual cues, only 11% of the words are correctly decoded. The two lateral slits, by themselves, provide only 19% intelligibility. Combining the two modalities results in nearly two-

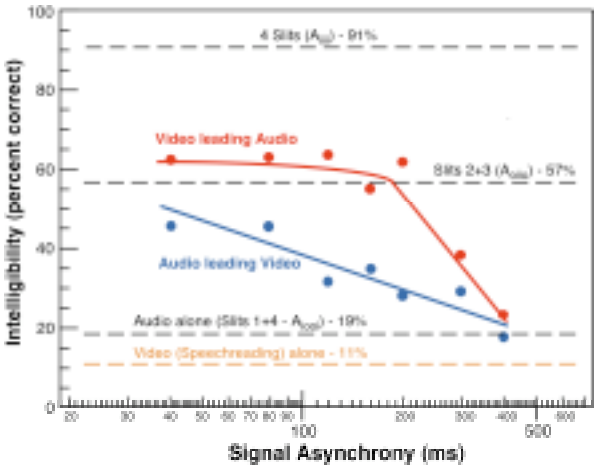

Figure 5: Average intelligibility (for 9 subjects) associated with audio-visual speech recognition as a function of bi-modal signal asynchrony. The audio-leading-video conditions are marked in blue, the video-leading-audio conditions shown in red. Baseline audio-only conditions are marked in black, dashed lines, and the video-alone condition is shown in orange. From [7].

thirds of the words (63%) being correctly identified – when the audio and visual streams are synchronized. As the two sensory streams are desynchronized an interesting phenomenon occurs. When the audio signal leads the video, intelligibility declines progressively in a manner similar to that observed for acoustic slit desynchronization shown in Figure 3 (Figure 5). However, when the video stream leads the audio, intelligibility remains unaffected until the magnitude of asynchrony is greater than 200 ms (Figure 5) [7]. For eight of the nine listeners participating in this experiment, the best performance observed was for conditions in which the video led the audio by intervals ranging between 80 and 120 ms.

There are many prospective explanations for the temporal asymmetry in audio-visual integration (see [10]). Perhaps the most compelling relates the visual modality to place-of-articulation information distributed over syllabic intervals. Concurrent or advance information pertaining to articulatory place provides a potentially important context with which to interpret the acoustic component of the speech signal.

7. The Relation of Time to Information

The temporal congruence of auditory, visual and motor mechanisms in speech is likely to reflect a common factor, one that is central to information processing in the brain. The syllable, ranging in length between 100 and 300 ms, is the linguistic manifestation of this interval shared in common across the sensory and motor systems, and is the most fundamental carrier of information in spoken language.

Prosodic phenomena, such as pitch and stress accent, are fundamental to all spoken languages and use the syllable as the basic medium for highlighting specific components of the message. Syllables vary in duration over a large range (see Figure 4). Much of this variation (at least in English) reflects prosody and the differential amount of information conveyed over the course of an utterance (see [12]). It is tempting to conclude that speech is spoken at rates varying between 3 and 7 syllables per second because this is the rate at which information contained in the speech signal can be reliably decoded by the listener, and in essence reflects the "sampling rate of consciousness."

8. References

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

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

[3] Dau, T., Kollmeier, B. and Kohlrausch, A. "Modeling auditory processing of amplitude modulation: II. Spectral and temporal integration," *J. Acoust. Soc. Am.* 102: 2906–2919, 1997.

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

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

[6] Dudley, H. "Remaking speech," *J. Acoust. Soc. Am.* 11: 169-177, 1939.

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

[8] Grant, K.W. and Greenberg, S. "Spectro-temporal interactions in auditory and auditory-visual speech processing," *Proc. 8th European Conf. Speech Comm. Tech. (Eurospeech-2003),* pp. 2557-2560, 2003.

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

[10] Greenberg, S. and Arai, T. "What are the essential cues for understanding spoken language?" *IEICE Trans. Inf. Sys.* 5, 2004 (in press).

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

[12] Greenberg, S., Carvey, H., Hitchcock, L. and Chang, S. "Temporal properties of spontaneous speech – a syllable-centric perspective," *J. Phonetics* 31: 465-485.

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

[14] Kozhevnikov, V.A. and Chistovich, L.A. *Speech: Articulation and Perception*. Joint Publications Research Service: Washington, D.C., 1966.

[15] Liberman, A.M., Cooper, F.S., Shankweiler, D. P., Studdert-Kennedy, M. "Perception of the speech code," *Psych. Rev.* 74: 431-461, 1967.

[16] Shannon, R.V., Zeng, F.G., Kamath V. and Wygonski J. "Speech recognition with primarily temporal cues," *Science* 270: 303-304, 1995.

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