

Time is of the Essence – Dynamic Approaches to Spoken Language

Steven Greenberg

The Speech Institute
Oakland, CA 94619 USA
steveng@cogsci.berkeley.edu

Abstract

Temporal dynamics provide a fruitful framework with which to examine the relation between information and spoken language. This paper serves as an introduction to the special Eurospeech session on “Time is of the Essence – Dynamic Approaches to Spoken Language,” providing historical and conceptual background germane to timing, as well as a discussion of its scientific and technological prospects. Dynamics is examined from the perspectives of perception, production, neurology, synthesis, recognition and coding, in an effort to define a prospective course for speech technology and research.

1. A Dynamic Approach to Spoken Language

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, known as 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. The special Eurospeech session on “Time is of the Essence – Dynamic Approaches to Spoken Language” is designed to acquaint the speech community with current research representative of this new emphasis on dynamics from a broad range of scientific and technical perspectives. The current paper serves as a brief introduction, providing historical and conceptual background for the session as a whole.

Traditionally, articulatory mechanisms 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 governed by vocal production, how does the brain decode speech 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 [17]. 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.

Motor Theory arose in response to the classical encoding problem in speech. Words are composed of constituent sounds, known as *phones*, which are represented in abstract form as a

linear sequence of *phonemes*. Somewhere in the brain resides a mental lexicon in which words are linked to their phonemic constituents. A central problem for models of spoken language is that phonemes are not packaged as discrete units in the acoustics or even in vocal production. The linguistic unit of closest affinity to the acoustics (and articulation) is the syllable not the phone [10]. A syllable contains anywhere from one to several phones, depending on its structure and context. Most syllables contain two or three phones, one of which is (almost always) a vowel, the other constituents serving as consonants. Through a process known as “co-articulation” vowels and consonants fuse within the syllable, certain features (e.g., “place of articulation”) spanning more than a single phone. Motor Theory treats this phonetic fusion as part of the encoding/decoding process.

A central issue for Motor Theory and other models of speech 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?

The papers in this special session address this fundamental issue from the perspective of dynamics. Three of the contributions focus on scientific approaches [9][19][20], while the others concentrate on technical applications [3][15][24].

A common theme 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 *very* slow variation in acoustic energy, spanning intervals between 40 and 400 ms [8]. 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 [8]. 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, spawned at Bell Laboratories, was initially taken most seriously by Chistovich and her colleagues in Leningrad, beginning in the 1960’s. They were perhaps the 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 [16].

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 [14]. The modulation spectrum’s peak in highly intelligible environ-

ments was found to be ca. 4 Hz, with a broad distribution of energy between 2 and 10 Hz (see Figure 1, lower panel). They correctly noted that 4 Hz conformed 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 becomes difficult to understand precisely under 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 assumptions directly, demonstrating that intelligibility does indeed depend on the integrity of the modulation spectrum below 8 Hz [6][7].

In Dudley's original conception 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 ca. 10 distinctive channels to produce good sounding speech [8]. In the mid-1990's Shannon and colleagues developed a form of VOCODER that questioned this parity between carrier and modulator [22]. In their study only four distinct channels were required to produce intelligible speech. Moreover, the carrier used was not harmonically structured as would be found in voiced speech, but rather Gaussian noise, akin to the glottal source associated with whispering [22]. 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 be largely dispensed with as long as certain essential modulation properties associated with the original signal are preserved [2][4][12] and their temporal (i.e., phase) relation across the acoustic frequency plane maintained [11][12][23] (see Figure 2). Intelligibility depends on both the *magnitude* and *phase* components of the modulation spectrum – in other words, the dynamics are key for understanding spoken language.

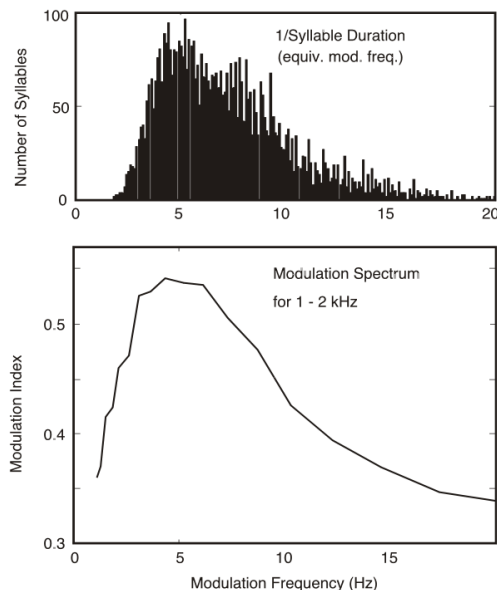


Figure 1. The relation between the distribution of syllable duration (transformed into modulation frequency) and the modulation spectrum of spontaneous Japanese material, computed for the octave region between 1 and 2 kHz. Adapted from [1].

2. The Eyes Have It

The first paper in the session, entitled “Spectro-temporal Interactions in Auditory and Auditory-Visual Speech Processing,” focuses on the role played by the dynamics of the visible articulators (“speechreading”) in understanding spoken language [9]. Grant and Greenberg emphasize the *complementary* nature of speechreading cues (i.e., the movement of the lips, jaw and tongue) relative to the audio signal – different parts of the frequency spectrum are associated with distinctive types of dynamics (see Figure 2 – “Waveform” panel for an example), which when combined, provide a high degree of intelligibility. Speechreading information is associated primarily with the

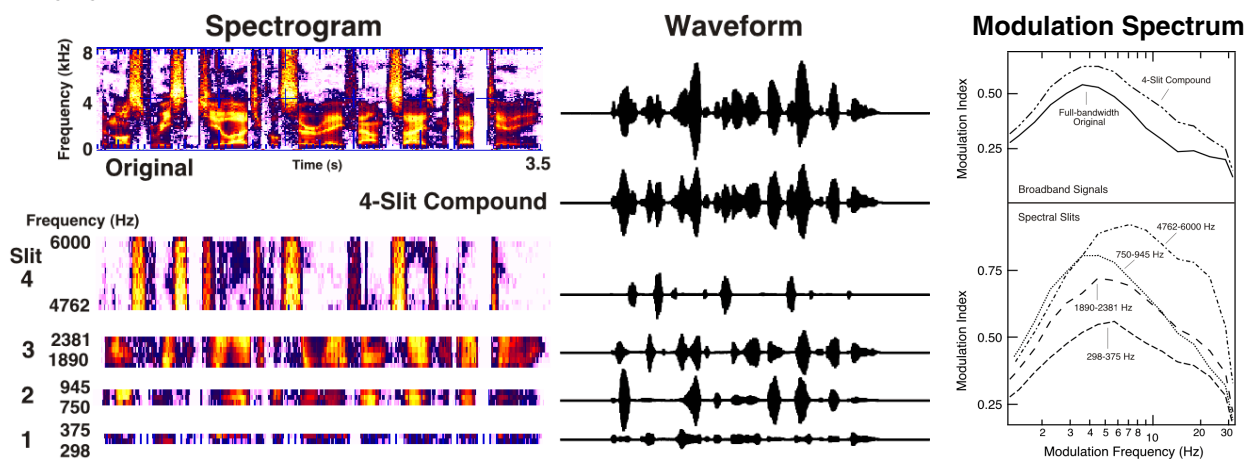


Figure 2. Spectrographic and time-domain representations of a representative sentence (“The most recent geological survey found seismic activity”) used in an intelligibility study [12]. The waveforms are plotted on the same amplitude scale, while the scale of the original, unfiltered signal is compressed by a factor of five for illustrative clarity. The frequency axis of the spectrographic display of the channels has been non-linearly compressed for illustrative tractability. Note the quasi-orthogonal temporal registration of the waveform modulation pattern across frequency channels. From [12]. In the right-hand panel is shown modulation spectra (magnitude component) associated with each of four, 1/3-octave channels, as computed for all 130 sentences presented in the experiment [bottom panel]. The peak of the spectrum (in all but the highest channel) lies between 4 and 6 Hz. Note the large amount of energy in the higher modulation frequencies associated with the highest frequency channel. The modulation spectra of the 4-channel compound and the original, unfiltered signal are illustrated for comparison [top panel]. From [12].

mid-frequency region of the spectrum, between ca. 800 and 3000 Hz, where place-of-articulation cues (e.g., the distinction between [p], [t] and [k] or [b], [d] and [g]) are concentrated. Such information can significantly improve speech comprehension in noisy environments (by ca. 8 dB in terms of S/N), particularly for the hearing impaired. Clearly, visible speech provides powerful (and effective) information concerning the mid-frequency dynamics that can be extremely effective in decoding the speech signal under conditions of extreme ambiguity. One possibility is that the modulation spectrum can be extracted from speechreading in a manner analogous to the acoustic signal.

How speech information from audition and vision is combined in the brain is unknown. The time frame over which dynamic information is fused appears to depend on which modality – vision or audition – encounters the signal first. If the audio signal leads the video, the time constraints for combining information are similar to those associated with audio-only conditions – there is a progressive decline in intelligibility for modality asynchronies as small as 40 ms. However, when the video stream leads the audio, intelligibility is relatively unaffected for asynchronies as long as 200 ms (i.e., the average duration of a syllable) [9]. It is as if the dynamics associated with speechreading are largely syllabic in nature, and that when the video signal arrives in advance of the audio, the time span germane to combining information from the two modalities is one adapted to syllable-length chunks.

3. Time Constants in the Brain

In his paper on “Brain Imaging Correlates of Temporal Quantization in Spoken Language” Poeppel examines the processing time associated with different components of speech as reflected in specific regions of the brain [19]. In particular, he shows that certain regions of the auditory system beyond primary auditory cortex exhibit a lateral asymmetry with respect to processing time constants – the left hemisphere (in right-handed subjects) tends to be more highly activated for signals containing rapid spectral transitions (ca. 40 ms), while the right hemisphere appears to be more involved in the processing of longer duration signals (ca. 200 ms). It is tempting to conclude from such studies that the left-side focuses more on mid- and high-modulation frequencies (> 16 Hz), while the right concentrates on the low-frequency modulation spectrum, particularly in the range of syllables (2-8 Hz). However, it is clear from Poeppel’s data that *both* hemispheres are activated during speech processing and that whatever differences are observed with respect to lateralization of function is one of degree rather than of kind [19].

4. Intrinsic Time Constants of the Vocal Apparatus

As mentioned in Section 1, the time constants characteristic of articulatory motion have traditionally been ascribed to biomechanical factors transcending speech. Saltzman, in his paper on “Temporal Aspects of Articulatory Control,” provides a much richer theoretical framework for understanding the temporal properties of spoken language [20]. In his view it is not only the velocity of the articulators that need to be considered, but also their acceleration and timing relative to each other (i.e., “phasing”). Many phonetic and prosodic phenomena appear to be accounted for within this framework, particularly when the concept of coupled oscillators is introduced within the model. One specific prediction made by Saltzman’s multi-tier framework is the “c-center” phenomenon in which consonant clusters at syllable onset act temporally as an integrated, tightly bound unit, whereas equivalent clusters in the coda behave more like individual segments. This result is of interest because

it provides an articulatory basis for the differential entropy associated with syllable onsets and codas. It has been observed that consonant clusters at syllable onset are much more likely to be pronounced canonically than their coda counterparts [10], implying that the onset contains more information (and greater stability) than the coda. Saltzman’s model is also consistent with the greater elasticity of consonant duration in the onset relative to the coda (as conditioned by syllable prominence) and provides a motivated mechanism for prosody’s impact on articulation [13].

5. Speech Timing – Natural and Synthetic

Models of articulation have traditionally served as the foundation for speech synthesis. However, as Zellner Keller points out in her paper on “The Temporal Organisation of Speech as Gauged by Speech Synthesis” timing has generally not been explicitly modeled in most of these systems. In her view prosodic properties are paramount for specifying both the temporal and phonetic characteristics of spoken language – therefore, it is essential to explicitly incorporate time into production models used for synthesis [24].

Although she does not use Saltzman’s model as the basis of her synthesis system, the approach taken by Zellner Keller lends itself well to the hierarchical framework described in his paper. Both authors believe that it is essential to model the production process from the perspective of multiple levels that interact with each other in temporally governed ways. In Zellner Keller’s view, the adequacy of the model can be most effectively evaluated by results of the synthesis, particularly the system’s capability to capture the fine temporal nuances associated with prosody, emotion and speaking style.

6. The Interaction of Time and Frequency

The challenge for automatic speech recognition differs from that of synthesis. In the latter, quality and naturalness are paramount, while for recognition it is the distribution of entropy, regardless of its form, that matters. In his paper on “Localized Spectro-Temporal Features for Automatic Speech Recognition” Kleinschmidt searches for the invariance in the speech signal associated with lexical identity [15]. He uses recent advances in auditory cortical physiology [5][18] as inspiration for utilizing Gabor functions that combine time and frequency into a single representation. In Kleinschmidt’s view, combining time and frequency into a joint representation provides greater potential for extracting truly meaningful cues in the acoustics than operating on time or frequency separately. This perspective is consistent with Grant’s studies on audio-visual integration [9] in that various regions of the acoustic spectrum appear to be associated with different aspects of temporal dynamics that combine in unique ways for decoding the speech signal.

7. A Mathematical Relationship Between Modulation and Information

The Gabor functions used by Kleinschmidt provide a convenient mathematical means with which to combine time and frequency into a single representation analogous to the processing performed by the auditory cortex in response to speech and other complex signals. In the paper by Atlas on “Modulation Spectral Filtering of Speech” the objective is somewhat different. Atlas is concerned with efficient coding strategies for speech and other audio signals [3]. Although he also uses auditory cortical processing as inspiration for his representational framework, he is ultimately interested not in linguistic units per se, but rather in information contained in the signal irrespective of its communicative significance (but defined in a more rigorous mathematical sense, e.g., [21]). To the extent

that entropy associated with different parts of the *modulation* spectrum distributed across the *acoustic* frequency spectrum can be mathematically characterized, it is possible to design extremely efficient coding algorithms capable of enormous compression with virtually no impact on intelligibility or sound quality. In Atlas' view, the modulation spectrum offers an excellent representation for efficient compression because of the wealth of perceptual and physiological support for its importance in the coding of speech and other acoustic signals.

8. What is the Essence of Time in Spoken Language?

Time is often 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, time 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?
- (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. This special session, "Time is of the Essence," will hopefully help us to better understand the essence of spoken language.

Acknowledgements

I thank Les Atlas, Ken Grant, Brigitte Zellner Keller, Michael Kleinschmidt, David Poeppel and Elliot Saltzman for participating in the Eurospeech special session on "Time is of the Essence – Dynamic Approaches to Spoken Language" and for taking the time to summarize their research in written form.

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] Atlas, L. "Modulation spectral filtering of speech," Proc. 8th European Conf. Speech Comm. Tech. (Eurospeech-2003), 2003.
- [4] 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.
- [5] Depireux, D.A., Simon, J.Z., Klein, D.J. and Shamma, S.A. "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," J. Neurophysiol. 85, pp. 1220–1234, 2001.
- [6] Drullman, R., Festen, J.M. and Plomp R. "Effect of temporal envelope smearing on speech reception." J. Acoust. Soc. Am. 95: 1053-1064, 1994.
- [7] 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.
- [8] Dudley, H. "Remaking speech," J. Acoust. Soc. Am. 11: 169-177, 1939.
- [9] 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), 2003.
- [10] Greenberg, S. "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation." Speech Comm. 29: 159-176, 1999.
- [11] Greenberg, S. and Arai, T. "The relation between speech intelligibility and the complex modulation spectrum." Proc. 7th Eur. Conf. Speech Comm. Tech. (Eurospeech-2001), pp. 473-476, 2001.
- [12] 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.
- [13] Greenberg, S., Carvey, H., Hitchcock, L. and Chang, S. "Beyond the phoneme – A juncture-accent model for spoken language." Proc. 2nd Int. Conf. Human Lang. Tech. Res., pp. 36-43, 2002.
- [14] 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.
- [15] Kleinschmidt, M. "Localized spectro-temporal features for automatic speech recognition," Proc. 8th European Conf. Speech Comm. Tech. (Eurospeech-2003), 2003.
- [16] Kozhevnikov, V.A. and Chistovich, L.A. *Speech: Articulation and Perception*. Washington, D.C.: Joint Publications Research Service, 1966.
- [17] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., Studdert-Kennedy, M. "Perception of the speech code," Psych. Rev. 74: 431-461, 1967.
- [18] Miller, L.M., Escabi, M.A., Read, H.L. and Schreiner, and C.E. "Spectrotemporal receptive fields in the lemniscal auditory cortex," J. Neurophysiol. 87: 516–527, 2002.
- [19] Poeppel, D. "Brain imaging correlates of temporal quantization in spoken language," Proc. 8th European Conf. Speech Comm. Tech. (Eurospeech-2003), 2003.
- [20] Saltzman, E. "Temporal aspects of articulatory control," Proc. 8th European Conf. Speech Comm. Tech. (Eurospeech-2003), 2003.
- [21] Shannon, C. and Weaver, W. *The Mathematical Theory of Communication*. Urbana: Univ. Illinois Press, 1949.
- [22] Shannon, R.V., Zeng, F.G., Kamath V. and Wygonski J. "Speech recognition with primarily temporal cues," Science 270: 303-304, 1995.
- [23] 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.
- [24] Zellner Keller, B. "The temporal organisation of speech as gauged by speech synthesis," Proc. 8th European Conf. Speech Comm. Tech. (Eurospeech-2003), 2003.