SOME PSYCHOACOUSTIC FACTORS IN PHONETIC ANALYSIS

Pierre L. Divenyi, Veterans Administration Medical Center, Martinez, California, 94553.

From an ethological point of view, speech represents a complex acoustic stimulus that has the greatest survival value for man. Physically speaking, speech is complex in two ways: its spectral composition, over any epoch of arbitrary length, is extremely rich, and this spectral composition is continuously varying over time. The information density represented by the speech signal is enormous; yet, the human auditory system, despite its limited capacity, is able to receive and decode such a complex signal with remarkable efficiency. The desire to provide a reasonable explanation for such efficiency, as well as the need for descriptive data on the perceptual processes that permit reception and decoding of speech, provided much of the motivation behind the greatest part of the speech perception research accomplished to date. The emerging body of experimental findings, in turn, has constituted the background for a number of theories and models of speech perception. The leitmotiv of many of these theories, including some major contemporary ones, is that speech represents a special acoustic signal that must be handled by the auditory system in a special way (=speech mode), involving special processes and mechanisms (=phonetic feature detectors, etc.). While the special nature of speech and speech perception processes can hardly be disputed (because of their aforementioned high survival value), some recent results demonstrating speech discrimination by young infants and animals have established the need for an alternative theoretical approach -- one that would take into account, at least to some extent, some "wired-in" properties of the auditory mechanisms. The purpose of the present paper is to invoke some basic properties of the human auditory system and to reflect on the consequences of these properties for the phonetic analysis of the speech signal.

Psychophysical reality of the speech signal

Classical psychoacoustics research and classical speech perception research have progressed on traditionally separate (and not always parallel) paths. The reasons for this divorce, considered by some cynics as permanent until quite recently, were numerous, one of them being the overwhelming concern of psychoacousticians with simple acoustic signals and peripheral auditory processes. However, for the last couple of decades, the situation has gradually changed: availability of sophisticated stimulus control, the growing popularity of a systems approach to perceptual problems, and interdisciplinary orientation of an increasing number of researchers have signaled the beginnings of a (hopefully) new era. Indeed, psychoacoustics appears to be no longer afraid of spectrally and/or temporally complex sound patterns and researchers seem to address with greater freedom issues involving more central processes. Thus, it has become possible to take a fresh look upon the speech signal as a stimulus to the auditory system, and to interpret its perception in terms of a certain number of discrete psychoacoustic processes. For reasons of economy, only a few major ones will be discussed here.

Peripheral analysis and time-frequency trade. Peripheral analysis of auditory signals operates under a constraint not unlike Heisenberg's Uncertainty Principle, as defined for elementary particle physics. According to this principle, in any given system frequency resolution (Δf) can be traded for temporal resolution (Δt) and vice versa, such that their product $\Delta f \Delta t$ remains constant. In the ear, such a relation is generally true only within certain limits (McGill, 1968); spectral resolution is limited by the Critical Bands (roughly 1/4 to 1/3 octaves in width; Zwicker et al., 1957) and temporal resolution by the ear's "time window" (a time constant of roughly 8 msec; Penner, 1978). Within these limits, however, this principle predicts that, to increase resolution in the spectral domain, temporal resolution must be sacrificed, and vice versa (Ronken, 1971). The validity of this prediction is proved by experimental results: discrimination of the frequency of pure tones deteriorates as their duration decreases (Moore, 1973) and, conversely, perception of the fine temporal structure of the stimulus is possible only for wide-band signals (Green, 1971).

Thus, the length of the effective time window and the width of the effective internal filter continuously adapt themselves to the spectral-temporal characteristics of the stimulus. The outcome of such an analysis will be a sequence of "neural spectra" (Klatt, 1978) or "central spectra" (de Boer, 1977) -- a series of quasi-stationary auditory events of variable duration. The temporal constraint signifies that peripheral analysis of acoustic (speech or non-speech) signals cannot be extended beyond the duration of these auditory events.

Pitch perception. According to contemporary theories (Plomp, 1975), pitch of complex signals is extracted by periodicity analysis of the internal spectrum (i.e., by taking its Fourier transform). Thus, any complex signal gives rise to two different pitch experiences: a "spectral pitch" (=formant analysis) and a "virtual pitch" (or low pitch or residue pitch [=periodicity analysis]), the former being a prerequisite for the latter. The existence region of virtual pitch is limited to pitch periods not shorter than about 2 msec (< 500 Hz); the degree of its salience is a composite function of the spectral region (formant region), the serial number and the relative intensity of the component harmonics, and the periodicity rate itself (Ritsma, 1962). In complex signals consisting of several consecutive harmonics virtual pitch is determined by the eight lowest harmonics, especially those around the third (Houtgast, 1974, 264), but, interestingly, the fundamental is not dominant.

Virtual pitch is not an absolute concept: it reflects a statistical approximation to a periodicity that derives from the ensemble of peaks in the internal spectrum (de Boer, 1977). It has also been proposed (Terhardt, 1974) that virtual pitch actually represents a Gestalt property of complex sounds -- a property that is as much a result of learning as that of purely sensory processes. Such a hypothesis helps account for some systematic pitch shift phenomena that are otherwise difficult to interpret.

Temporal organization. Since peripheral analysis is limited to short temporal intervals, the sequences of "neural spectra" which temporally-complex signals generate must be organized into perceptually meaningful units by some higher-level auditory center(s). Such a perceptual organization in time obeys rules that are reminiscent of the Gestalt principles that govern the perception of visual figures in space (e.g., law of closure, law of proximity, etc; Koffka, 1935) and, ultimately, leads to the percept of an auditory pattern (Divenyi and Hirsh, 1978). Among the general rules of auditory pattern perception there is one of primary importance: two successive auditory events can be optimally resolved in time only if they occur in identical spectral bands. For example, auditory discrimination of short (10-30 msec) intervals defined by the onsets of two brief tones gradually deteriorates when the two tone frequencies become increasingly different (Divenyi and Sachs, 1978). Similarly, recognition of the temporal order of successive tones remains accurate only as long as all tone frequencies are within the same narrow band -- otherwise the sequence breaks into separate "auditory streams" (Bregman and Campbell, 1971).

The concept of "listening bands". The three above mentioned limitations, i.e., trade-off of time resolution - frequency resolution, limits of periodicity analysis, and restriction of accurate temporal organization to auditory events within the same narrow spectral band, are generally valid for the processing of any auditory signal, simple or complex. However, since speech constitutes an auditory stimulus in which the spectral information is generally distributed over several bands (specific to a given phonetic unit), its processing will be further complicated by yet another limitation: the auditory system is unable to simultaneously monitor several bands without loss of information (Green, 1961). The consequence of such a limitation is that auditory processing along various acoustic dimensions will be degraded by *frequency uncertainty*, i.e., by leaving the listener in doubt as to the frequency region in which the forthcoming auditory event is to appear. For example, frequency uncertainty will degrade detection (Creelman, 1972) and frequency discrimination (Watson, 1976) of a pure tone, as well as recognition of temporal-order patterns of several successive tones (Divenyi and Hirsh, 1978).

In order to overcome the effect of frequency uncertainty, the auditory system tends to spontaneously "*tune*" its focus of listening to the narrow band at or around the input frequency; it will usually *remain* focused at this listening band *in the absence* of any stimulus for at least several seconds (Johnson, 1978). Thus, at any given time, the auditory system's choice of a listening band is determined by the frequency characteristics of the *last input*. One of the possible reasons for the detrimental effect of frequency uncertainty is that shifting the listening focus from one band to another seems to *take time* (Divenyi and Hirsh, 1972). Moreover, attending to more than one spectral band at once will also degrade listening efficiency (Swets, 1963) -- the information processing capacity of the ear is, indeed, quite limited. The surprising finding is that the listener's knowledge with regard to the frequency of the forthcoming stimulus is *not* sufficient to completely eliminate the frequency uncertainty effect: to tune the listening band to a new region some sound (i.e., a *cue*) must occur (Johnson, 1978).

The *locus* of the tuning mechanism most probably lies above the auditory periphery: contralateral cues, too, have been found to be effective in establishing the listening bands (Gilliom et al.,1979)

Relevance to speech perception

The question of great interest to many is how a system having the properties described above is likely to behave when confronted with a speech signal. While a great deal more experimental data than what we have to date are needed to answer this question (even in a marginally acceptable manner), it is nonetheless possible to give a cursory outline of the effects of auditory processing on speech sounds. Again, because of space limitations, the picture presented here will be sketchy and less than exhaustive.

Segmentation. As a direct result of the time resolution - frequency resolution trade-off, any complex signal in which narrow-band and wide-band portions alternate will be automatically segmented at a peripheral level. Since, in speech, transitions from wide-band to narrow-band acoustic segments (and vice versa) roughly correspond to phonetic segment dividers, each of these transitions (smoothed by the ear's time window function) will produce *marker signals* at the auditory periphery. Thus, the series of auditory events (="neural spectra") which some higher-level centers will organize into perceptual units will actually be a succession of phonetically meaningful elements.

Speaker invariance. The mutual interdependence of waveform periodicity, spectrum of complex sounds, salience of virtual pitch, and salience of spectral pitch can account for much of the formant frequency - fundamental frequency relations observed in vowel production and perception (Fujisaki and Kawashima, 1968). Since vowels (=quasi-steady-state sounds) are analyzed in a narrow-band mode, relatively small spectral variations may be detected by the auditory system. Such a large degree of sensitivity may provide the explanation underlying the notion that vowel perception is "continuous" rather than "categorical".

Categorical perception and selective adaptation. In CV syllables, especially in stop-vowel pairs, the initial consonant is a wide-band transient; therefore, nothing compels the auditory system to tune the listening band to any particular position of the spectrum. The relative freedom of tuning that derives from wide-band stimuli enables the auditory system to select a frequency region to which it will spontaneously direct its focus *before* the onset of the CV sound. Such strategies may possibly originate in learning:

category boundaries that characterize certain features are known to be language-bound. However, strategies for positioning the listening band are by no means absolute: a sound of different spectral-temporal characteristics (speech or non-speech, see Samuel and Newport, 1979) presented prior to the CV stimulus could serve as a cue (Johnson, 1978) and make the auditory system choose a different listening band. Thus, selective adaptation effects could be re-interpreted in terms of pre-cueing and listening bands.

Such an interpretation is quite straightforward when one looks upon category boundary shifts observed for the feature of place-of-articulation in adaptation experiments: the acoustic basis for this feature is almost exclusively spectral. Explanation of boundary shifts of the voiced-voiceless category, a predominantly temporal feature, is somewhat more complex. Since temporal organization of acoustic events heavily depends on temporal cues contained in some narrow band, perception of the feature of voicing will be a function of the discriminability of voice-onset-time inside one (or several) narrow spectral region(s). However, when a brief auditory time interval is marked by a pair of sounds of identical spectral composition, *temporal masking* (forward or backward) of one marker by the other could decrease the discriminability of the interval (Divenyi and Sachs, 1978). Because the relative energy of the consonant and the vowel varies from one band to another (thereby also causing the amount of temporal masking to vary), the choice of the monitored band will be critical in determining the VOT boundary. Thus, tuning the listening band to different spectral regions will result in different voicing boundaries. An adaptor stimulus (by virtue of its potential role as a cue), therefore, may alter the natural position of the listening band for a given CV syllable, thereby producing a shift in the category boundary. It is conceivable that perceptual-productive acquisition of different phonetic patterns could also be associated with different spectral positions that the listening band will spontaneously occupy; thus, the present theory is consistent with the language-dependent nature of voicing category boundaries.

Time invariance implies that the relative duration of certain phonetic segments is irrelevant. Experiments on the perception of non-speech sound sequences (Watson, 1976; Divenyi and Hirsh, 1978) have shown that the emergence of an auditory pattern (at least within certain limits) *does not depend* on the absolute duration of the components. Thus, it follows that the rate at which the speech segments ("neural spectra") of the speech sounds occur will not change the "figural properties" of the patterns.

Conclusion: Whither phonetic analysis?

When attempting to examine speech processing on the auditory level, one finds that the product of auditory analysis possesses several characteristics that are customarily thought to belong to the realm of phonetic analysis (feature analysis, etc.). While it is readily acknowledged here that many crucial experiments needed to prove (or disprove) critical points have not yet been performed, and that straight extrapolation of non-speech auditory data to speech-bound processes may often be risky, we feel, nevertheless, that auditory analysis of the speech signal well exceeds the limits imposed on it by several widely accepted theories. The view that phonetic analysis may not be an indispensable stage in speech processing is concordant with the opinion expressed in some studies on the perception of speech by man (Ades, 1976) or the recognition of speech by machine (Klatt, 1978). An alternative view, one that we would like to propose herewith, is that speech perception may be regarded as a special class of auditory pattern perception -- *special* only because we have learned these patterns so well.

References

Ades, A.E. (1976): "Adapting the property detectors for speech perception", in *New approaches to language mechanisms*, R.J. Wales and E. Walker (eds.), 55-1o8, Amsterdam: North Holland.

de Boer, E. (1977): "Pitch theories unified", in *Psychophysics and physiology of hearing*, E.F. Evans and J.P. Wilson (eds), 323-334, London: Academic.

Bregman, A.S. and J.L. Campbell (1971): "Primary auditory stream segregation and perception of order in rapid sequences of tones", *JEP* 89, 244-249.

Creelman, C.D. (1972): "Detecting signals of uncertain frequency: Analysis by individual alternative signals", *JASA* 52, 167.

Divenyi, P.L. and I.J. Hirsh (1972): "Discrimination of the silent gap in two-tone sequences of different frequencies", *JASA* 51, 138.

Divenyi, P.L. and I.J. Hirsh (1978): "Some figural properties of auditory patterns", *JASA* 64, 1369-1385.

Divenyi, P.L. and R.M. Sachs (1978): "Discrimination of time intervals bounded by tone bursts", *Perc. Psych*. 24, 429-436.

Fujisaki, H. and T. Kawashima (1968): "The roles of pitch and higher formants in the perception of vowels", *IEEE AEA* AU-16, 73-77.

Gilliom, J., D.W. Taylor and C. Cline (1979): "Timing constraints for effective cueing in the detection of sinusoids of uncertain frequency", Perc. Psych. 25 (in press).

Green, D.M. (1961): "Detection of auditory sinusoids of uncertain frequency", JASA 33, 897-903.

Green, D.M. (1971): "Temporal auditory acuity", Psych. Rev. 78, 540-551.

Houtgast, T. (1974): "Masking patterns and lateral inhibition", in Facts and models in hearing, E. Zwicker and E. Terhardt (eds.), 258-265, Berlin: Springer.

Johnson, D.M. (1978): "Attentional factors in the detection of uncertain auditory signals", Unpubl. Doct. Dissert. Univ. Calif. Berkeley.

Klatt, D.H. (1978): "Speech perception: A model of acoustic-phonetic analysis and lexical access", in Perception and production of fluent speech, R.A. Cole (ed), Hillsdale (N.J.): Erlbaum.

Koffka, K. (1935): Principles of Gestalt psychology, New York: Harcourt Brace.

McGill, W.J. (1968): "Polynomial psychometric functions in audition", J. Math. Psych. 5, 369-376.

Moore, B.C.J. (1973): "Frequency difference limens for short-duration tones", JASA 54, 610-619.

Penner, M.J. (1978): "A power-law transformation resulting in a class of short-term integrators that produce time-intensity trades for noise bursts", JASA 63, 195-201.

Plomp, R. (1975): "Auditory psychophysics", Ann. Rev. Psych. 26, 207-232.

Ritsma, R.J. (1962): "Existence region of the tonal residue", JASA 34, 1224-1229.

Ronken, D.A. (1971): "Some effects of bandwidth-duration constraints on frequency discrimination", JASA 49, 1232-1242.

Samuel, A.G. and E.L. Newport (1979): "Adaptation of speech by nonspeech: Evidence for complex acoustic cue detectors", JEP HPP 5 (in press).

Swets, J.A. (1963): "Central factors in auditory frequency selectivity", Psych. Bull. 60, 429-440.

Terhardt, E. (1974): "Pitch, consonance, and harmony", JASA 55, 1061-1069.

Watson, C.S. (1976): "Factors in the discrimination of word-length auditory patterns", in Hearing and Davis: Essays honoring Hallowell Davis; S.K. Hirsh, D.E. Eldredge, I.J. Hirsh, and S.R. Siverman (eds.), 175-189, St. Louis: Washington University Press.

Zwicker, E., G. Flottorp and S. S. Stevens (1957). "Critical band width in loudness summation", JASA 29, 548-557.