COCHLEAR IMPLANT
Acquisitions and Controversies

edited by

BERNARD FRAYSSE, M.D.

and

NADINE COCHARD

With the collaboration of:

Françoise DELAVIE, Olivier DEGUINE, Marie-Josée ESTEVE-FRAYSSE,
François FURIA, Marie-Laurence LABORDE, Françoise SONILHAC,
Henri URGELL, Geoffroy VANDEVENTER, Marie-Anne VINCENT

Temporal information in speech and its relevance for Cochlear implants

STUART ROSEN, Ph. D.

ABSTRACT

The auditory capabilities of users of cochlear implants are strongly dependent upon the temporal features of auditory stimulation. A new framework for describing the acoustic structure of speech based purely on temporal properties has been developed to clarify these abilities. From this point of view, speech can be said to be comprised of three main temporal features (based on dominant temporal frequencies): (1) envelope (2) periodicity, and (3) fine-structure. The various types of segmental and prosodic linguistic information signalled by each feature is described, and the extent to which the salience of each varies across different groups of listeners discussed. A pilot experiment with normal listeners, in which the temporal structure of speech is preserved while eliminating spectral structure, provides empirical support for this three-way system of temporal features. Further clarification of the role of temporal features in audition will lead not only to theoretical advances in understanding both electrical and normal hearing, but also practical benefits in efficient utilisation of limited electro-auditory capacity in users of cochlear implants.

INTRODUCTION

One of the most long-standing controversies in hearing theory concerns the extent to which place- and/or time-based features are responsible for such basic auditory percepts as melodic pitch and timbre. There are two main reasons for current interest in the role of temporal information in the perception of complex auditory signals, particularly speech. Firstly, theoretical models derived from physiological evidence (Sachs & Miller, 1985; Sachs, Young & Miller, 1983) suggest not only that temporal information might be important for the perception of melodic pitch, but that it is also implicated in the perception of spectral shape variations (one important aspect of timbre). Secondly, a large number of patients have received cochlear implants that deliver a signal based on the speech waveform to a single electrode, thus allowing no place-based frequency analysis. Yet, many of these patients have performed surprisingly well, even to the extent of being able to understand unknown sentences on

the basis of an auditory signal alone (Hochmair-Desoyer et al., 1980, 1985). These results have stimulated hypotheses about the extent to which "temporal" information on its own can be effective in speech perception.

A FRAMEWORK FOR DESCRIBING TEMPORAL INFORMATION IN SPEECH

There is, however, much confusion about the linguistic information contained in the temporal structure of speech, and the extent to which it is useful to users of cochlear implants. As a complement to the standard Fourier-based spectral approach (which is totally inapplicable to sensations derived from single-electrodes), a three-way classification of the temporal structure of speech is proposed, based on dominant temporal frequencies:

(1) "Amplitude envelope", "time/amplitude", or "time-intensity" information, fluctuations at rates between about 2 and 50 Hz in overall amplitude. We will refer to this simply as **envelope**. In much of the literature, this is what authors mean by "temporal information". Such low frequency variations in overall amplitude can convey four main types of linguistic information:

(a) Segmental cues to manner of articulation, in a wide variety of ways. Consider, for example, the affricate/fricative distinction contrasting the English words "chip" and "ship", for which a number of envelope features are known to be influential (Dorman, Raphael & Isenberg, 1980; Gerstman, 1957; Howell & Rosen, 1983, 1987; Repp et al., 1978). The frication noise in "ch" has a quicker rise time and shorter overall duration than the corresponding frication in "sh". "ch" typically has a short release burst whereas "sh" does not. Short release bursts typically indicate plosive type sounds. Silent gaps too may indicate the presence of a voiceless plosive (Bailey & Summerfiled, 1980; Summerfield, Bailey, Seton & Dorman, 1981). More generally, it has been proposed that relatively fast changes in overall amplitude mark consonants from non-consonants (Stevens, 1980, 1981; Stevens & Blumstein, 1981), or continuants from non-continuants (Shinn & Blumstein, 1984).

(b) Relatively weak segmental cues to voicing in certain segment types. Generally speaking, voiced sonorants (vowels, semivowels, nasals and laterals, for example, /m/, /l/) have a greater amplitude than voiceless obstruents (for example, /f/, /p/). The duration of silent intervals may be important in distinguishing voiced from voiceless plosives in inter-vocalic position. In some contexts, vowel duration, in so far as it is cued by envelope, can give information about voicing in the following consonant (Umeda, 1975).

(c) Weak segmental cues to vowel identity. The duration of vowels varies lawfully with vowel quality, and so can signal some information

about it. Many languages use duration contrastively (along with changes in quality) to distinguish among vowels (see Lehiste, 1970, pp. 18-19, 30-35 for a review). For example, all other things being equal, the vowel in "heed" tends to be of significantly longer duration than that of "hid".

(d) Prosodic cues. Dynamic envelope cues can be used to assist syllabification (as Mermelstein, 1975 has shown in an automated procedure), and relative amplitude (on a more static basis) probably plays a minor role in the assignment of stress in words (for example in distinguishing the verb "per**mit**" from the noun "**per**mit" – see Crystal, 1969, pp. 113-120; Fry, 1968; Lehiste, 1970, pp. 36-38, 120-139 for reviews of the relevant literature). In so far as amplitude onsets and offsets can demarcate linguistic units (vowel, syllable or word), much information about duration, and hence speech rhythm and tempo can also be extracted from envelope cues. Duration itself appears to play a role in word-level stress (see above) while information about tempo could assist listeners in normalizing for speech rate variations in segmental (Miller, 1981) and prosodic contrasts. Variations in speech rate can also carry distinctions in meaning (Crystal, 1969, pp. 152-156)[1] or indicate parenthetical comments.

(2) Periodicity or aperiodicity of stimulation, and the rate of periodic stimulation. Periodicity exists primarily in the region between about 50 and 500 Hz, while aperiodicity typically extends to the 5-10 kHz region. We will refer to these jointly as **periodicity** information. Because periodic and aperiodic sounds can differ so greatly in their frequencies, it may be useful at times to think of periodicity information as being divided into these two subclasses with different frequency content, but which both give information about the source of excitation in speech production. Periodicity information, in the more general sense, directly conveys two main types of linguistic information:

(a) Segmental information about voicing and manner. The presence of low-frequency quasi-periodic acoustic energy in a speech signal is a reflection of the quasi-periodic vibrations of the vocal folds. Such sounds are said to be voiced, and such voicing is the most important cue to the phonological feature of voicing, perhaps the most basic distinction in all of the world's languages. Similarly, in many languages (such as English and French) there is an association between manner and voicing features (all nasals are voiced) which manner information to be obtained from information about phonetic voicing patterns. Speech segments

1. Crystal (p. 153) gives the following example of the way in which variations in speech rate can carry distinctions in meaning. Consider the phrase "Those who aren't Christians, aren't Catholics ...". When uttered with an appropriate pause after "Christians" and normal tempo following, a type of clause coordination is indicated. The speaker could have gone on to say: "aren't Muslims, aren't Jews", etc. However, with little or no pause after "Christians", and increased tempo following, the speaker indicates s/he meant to say "Catholics" rather than "Christians". This is what Crystal deems as "the distinction between (one type of) coordination and what one might loosely call a 'slip of the tongue' ...".

which are aperiodic result from turbulence noise generated by aerodynamic flow between closely spaced articulators. Such aperiodicity can be a strong cue for voicelessness, and/or to the fricative manner of articulation.

(b) Prosodic information relating to intonation and stress. The fundamental frequency of quasi-periodic energy in a speech signal is a reflection of the vocal fold vibration rate, and is the prime acoustic correlate of the perception of voice pitch. Linguistically meaningful patterns of voice pitch are known as intonation and tone, and play important roles in accenting syllables in words and sentences (e.g., for emphasis), in clarifying ambiguous pronoun references, in marking syntactic units and in distinguishing questions and statements (for reviews see Fry, 1968; Lehiste, 1970; Rosen & Fourcin, 1986). Furthermore, in tone languages like Chinese, voice pitch patterns have a lexical function - that is, they distinguish different dictionary meanings of a word. For example, in Cantonese Chinese, the syllable "yee" may mean "clothes", "chair", "meaning", "child", "ear" or "two", depending upon the pitch contour used when uttering it. Even English has a minor instance of this, in that voice pitch contours can play an important role in distinguishing between the verbal and nominal function of a word (as in "per**mit**" vs. "**per**mit" - see above).

(3) Temporal fine-structure, variations of wave shape within single periods of voiced sounds, or over short time intervals of voiceless ones. This cue has dominant frequency components from about 600 Hz upward to about 10 kHz. We will call this **fine-structure** information. Acoustic fine-structure is strongly related to (but not completely determined by) the spectral content of a sound - its frequency spectrum[2]. It can convey at least two main types of linguistic information, both segmental:

(a) Segmental cues to voicing and manner. Voiced sounds have a spectrum heavily weighted to low frequencies (below about 500 Hz) whereas voiceless sounds typically have their peak energies at frequencies considerably higher. First formant transitions are known to play some role in distinguishing English voiced from voiceless plosives in initial prevocalic position (e.g., Hazan, 1989; Soli, 1983; Stevens & Klatt, 1974). Apart from the information signalled by voicing, other cues to manner may be signalled by the shape of the spectrum. Nasals, for example, are characterized by a low first formant frequency, broad resonances, and zeros in the spectrum (Fujumura, 1962). Stevens (1980, 1981) has discussed the role of sudden spectral changes (usually in conjunction with sudden envelope changes) in distinguishing consonantal sounds from non-consonantal ones.

---

2. Fine-structure depends on the phase spectrum of a sound, as well as its amplitude spectrum.

(c) Segmental cues to place of articulation and vowel quality. This function of fine-structure is by far the most important, not least because spectral shape variations are more or less the only acoustic cues to place. For example, it is well known that the two most important acoustic features that distinguish the syllables "ba", "da" and "ga" from one another are the frequency spectrum of the initial release burst, and the dynamic formant transitions that follow (e.g., Hazan, 1989 and references therein). Similarly, English voiceless fricatives in a prevocalic position may be distinguished from one another on the basis of static spectral shape, or the formant transitions in the following vowel, with the importance of each cue strongly dependent on the particular place of articulation (Harris, 1958). Finally, spectral shape is the major cue to vowel identity.

This list is not exhaustive, as there do seem to be weak cues to speech contrasts in temporal features that are not mentioned above[3]. However, the most important ways in which various temporal features figure in speech contrasts have been included, and they are summarized in Table 1. Note too that these features may not always operate independently of one another. Duration, which has been grouped as an envelope cue, usually refers to the duration of an interval of speech with particular acoustic properties. For example, it is the duration of *aperiodicity* that helps to distinguish "ch" from "sh". Finally, there is likely to be much overlap in the frequency region over which the features operate. The release burst which is present in "ch" but not in "sh" is so short that its envelope would certainly contain frequencies above 50 Hz. It is still the case, however, that the properties of the burst can be reasonably well divided among the features of envelope, periodicity and fine-structure.

FOR WHAT LISTENERS DO TEMPORAL FEATURES OF SPEECH OPERATE ?

When we come to consider the role of temporal features in speech perception, further complications arise depending upon the type of listeners we are dealing with. For users of single-channel implants, no frequency/place mechanisms operate and so *all* auditory perception must be based on temporal features. Here, the three-way framework detailed above will apply completely, and the only extra problem that need concern us is the extent to which temporal features are modified by the patient's speech processor. This can range from the relatively innocuous

3. For example, fine structure may contain weak prosodic cues. At least in English, the vowels in unstressed syllables tend to be more neutral in quality than the vowels in stressed syllables. There is some evidence that this feature influences decisions of the "**per**mit" vs. "**per**mit" variety (see above). So, too, may envelope give weak information about place of articulation, because the duration of aperiodicity in voiceless initial English plosives is known to vary lawfully with place of articulation (Lisker & Abramson, 1964).

automatic gain control of the Vienna device (which has little effect on any of the features) to the hard-clipping of the House/3M device (which has drastic effects on envelope and fine-structure, leaving perhaps only periodicity unaffected). These modifications of temporal features may be advantageous, of course. There is evidence, for instance, that the House/3M device may make certain envelope features (e.g., plosive release bursts) more salient (Rosen et al., 1989). Ignoring the details of these complications for the moment, it appears that most users of analogue single-channel implants are sensitive to envelope and periodicity, but the relative importance of each cue is not yet clear (Agelfors & Risberg, 1989; Rosen & Ball, 1986; Rosen et al., 1989, Tyler et al., 1987). Also, some patients may only be sensitive to the low-frequency periodicity of voiced sounds, missing out voiceless sounds completely (Rosen & Ball, 1986). In many, if not most, patients, there is also sensitivity to temporal fine-structure, but this is relatively rarely used in the perception of natural speech (Agelfors & Risberg 1987, 1989; Hochmair-Desoyer et al., 1985; Rosen & Ball, 1986; Rosen et al., 1989; White, 1983). However, the reception of unknown sentences by auditory means alone in users of single-channel implants (Hochmair-Desoyer et al., 1980, 1985) implies some linguistic use of the temporal fine structure of speech, although this hypothesis has yet to be explicitly confirmed.

At the other end of the observer continuum lie normal listeners, in whom the effects of peripheral auditory filtering must be considered[4]. Looking at a speech wave on an oscilloscope, we see a unitary trace of amplitude variations in time. The normal auditory system breaks this down, via the filtering action of the cochlea, into many waveforms, each of which will have its own three-way complement of temporal information. So, for instance, the envelope features transmitted by the auditory nerve

4. Plomp (1983) has described a three-way partition of the properties of speech sounds based on the concept of modulation, which bears much resemblance to the system described here. In Plomp's words: "Speech can be considered to be a wideband complex signal modulated continuously in time in three different respects: (1) the vibration frequency of the vocal cords is modulated, determining the pitch variations of the voice, (2) the temporal envelope of this signal is modulated by narrowing and widening the vocal tract locally by means of the tongue and lips, and (3) the tongue and the lips in combination with the cavities of the vocal tract determine the sound spectrum of the speech signal, which may be considered as a modulation along the frequency scale." These are clearly related to the periodicity, envelope and fine-structure categories described above. There are, however, a number of difficulties in this characterization. (1) Fine structure is only discussed via the frequency domain (i.e., as a spectrum). Although reasonable for a system aimed at explaining the perception of normal listeners, it limits its usefulness as regards implants and theories of normal hearing which require use of temporal features in the perception of spectral shape variations. (2) No mention is made of the existence of aperiodic sounds. (3) Aspects of envelope are controlled by vocal fold behaviour, and not by supralaryngeal manœuvres (e.g., the decrease in amplitude over the vowel for any CV syllable uttered in citation form, as in "key"). (4) The sound spectrum of voiced speech sounds is influenced by the spectrum of the source (determined by vocal fold behaviour), as well as by the shape of the vocal tract. Generally speaking, it appears that a description in terms of acoustic properties is more useful than a description in terms of production, not least because it eliminates descriptive difficulties like the last two just mentioned. In terms of traditional descriptions of speech production, only the temporal feature of periodicity has a simple productive correlate - that of the **source** of excitation. Fine structure results, as implied above, from the interaction between source and filter, as indeed does envelope.

will be modifications (to a greater or lesser extent) of those observed on a speech waveform. More importantly, peripheral auditory filtering means that temporal cues will be transformed into place cues, at least for periodicity and fine-structure. Although there is strong evidence that temporal features operate for these latter features even in the normal listener, no one doubts that peripheral place/frequency analysis plays a crucial role. Only for envelope features (and not totally even then) does it appear that purely temporal processes are active. One way to test the usefulness to normal listeners of purely temporal information in speech is to manipulate signals so as to eliminate spectral variations while retaining *only* temporal information. This may also be thought of as simulating a single-channel implant user with a normal listener.

Somewhere in the middle of this continuum on which listeners vary in the extent of their frequency/place analysis are users of multi-channel implants. In implants based on a filter-bank model of the auditory periphery (often mistakenly referred to as a "vocoder" scheme[5]), the spectral analysis of the normal ear is replaced, at least roughly, by electronic filters (as in the Symbion and San Francisco devices; Eddington, 1983; Merzenich, 1985). Again, we have the problem of disentangling place and time information, but one which is easily avoided by stimulating one electrode at a time, or all electrodes with the same signal. Even with multi-channel stimulation, the degree of frequency selectivity will be much less than in the normal ear, and so the role of explicitly temporal factors will be much larger than for normal listeners. For all implants, the degree of selectivity is likely to be much less than that indicated simply by considering the filtering properties of the speech processor, due to current spread in the cochlea. In some implants, only a very gross filtering (e.g., 4 channels in the Symbion device) is even attempted. It is therefore not surprising to find suggestions in the literature that high performance on a vowel identification task by Symbion users can be attributed primarily to the use of temporal fine-structure (at least in the first formant region, Dorman et al., 1989).

Multi-channel implants of the feature extracting sort (e.g., the Nucleus device, Seligman, 1987) demand a slightly different approach, in that there is usually an explicit design choice to signal certain acoustic features

5. A vocoder is an electronic device originally used to reduce the bandwidth needed for the transmission of speech without reducing its intelligibility much. Vocoders make use of a filter bank to analyze a speech signal, rectifying and smoothing its outputs so as to obtain an estimate of the energy within each filter passband. A separate determination of the periodicity or aperiodicity of the signal is made, and if the signal is periodic, the fundamental frequency is measured. These extracted parameters (aperiodic vs. periodic, and its fundamental if the latter, and the level of the spectrum for each filter in the bank) are then transmitted and used to resynthesize the speech at the receiving end. Cochlear implants like the San Francisco and Symbion devices use a bank of filters, and feed their outputs more or less directly to the set of electrodes. From the viewpoint of temporal complexity, a vocoder eliminates the temporal detail from the output of the filter bank, while the previously mentioned implant systems maintain it. In this respect, there is a strong argument that the Nucleus device is more like a vocoder than the other two, in that it too tries to determine the shape of the spectrum and eliminate temporal detail.

via temporal information, and others via differential places of stimulation. Theoretically, understanding patient performance should be more straightforward because the range of electrical patterns presented is simplified. In practice, so little is known about the actual behaviour of the speech extraction circuits (as opposed to their planned behaviour) that any effort saved by restricting the type of information presented to the patient must be expended towards characterizing the behaviour of the processor. A good example of this is found in comparing performance with two versions of the Nucleus device with regard to the transmission of voicing information for consonants in an intervocalic context. The newer version of the device uses extracted voice fundamental frequency to govern the rate of stimulation, while two estimates of energy concentration (in the first and second formant region) determine which two electrodes are stimulated per fundamental period. Patients using this device can show good sensitivity to voicing, unlike those using an earlier version of the device in which only one electrode relating to energy in the second formant region was stimulated at a rate determined by speech fundamental frequency (Blamey et al., 1987; Dowell et al., 1982, 1985, 1987). Why should this happen in a device which explicitly claims to extract fundamental frequency and present it in a temporal code which we know (both theoretically and practically - Rosen & Ball, 1986) to be perfectly adequate for transmitting voicing information? As noted above, information about voicing can come either from periodicity or spectral shape information. As only users of the later device, which can effectively signal spectral balance, do well in perceiving voicing, it may be that the temporal correlates of voicing are *not* well presented. In fact, there is evidence that the fundamental frequency extracting circuit of the Nucleus device does not work very well (Howard & Seligman, 1983).

Clearly, research on understanding implant patient performance with speech signals is in its infancy. The framework presented above is a useful starting point, but much remains to be done. For example, it is not yet clear the extent to which even normal listeners can use purely temporal information in the linguistic contrasts where it is at least theoretically useful. There is some evidence that normal listeners can use envelope and periodicity cues, even when they are purely temporal, but the role of finestructure is much more uncertain. One way to approach this question is (as mentioned above) to transform speech signals so that they contain temporal information only, and to present these to normal listeners. This approach has been used quite successfully in the psychoacoustic domain to show that the perception of melodic pitch can be based purely on temporal information (Burns & Viemeister, 1976). In that study, white noise was amplitude-modulated at various rates, and presented to listeners in a musical interval identification task. White noise has of course, a flat spectrum, and amplitude modulating it by any signal leaves the spectrum of the resultant signal flat as well. Therefore, these stimuli contained only temporal information. The fact that listeners could perform a musical interval identification task was

used to argue that the perception of musical pitch could be based on temporal cues alone.

TEMPORAL INFORMATION IN THE IDENTIFICATION OF CONSONANTS

Van Tasell et al. (1987) adopted a similar approach in an attempt to investigate the role of temporal information on a purely segmental level in consonant identification. They presented to normal listeners a set of 19 vowel-consonant-vowel (VCV) utterances that had been transformed so as to (supposedly) contain only envelope information. Three sets of stimuli were created by low-pass filtering full-wave rectified speech at cut-off frequencies of 20, 200 and 2000 Hz, and using the extracted "envelopes" to amplitude-modulate a pink noise.

This study has, however, important limitations, both empirical and conceptual. **(1)** Only a single token of each VCV was used, which may contain artefactual or unrepresentative features. This may explain unusual confusions in Van Tasell et al.'s data (e.g., in the low-pass 2000 Hz condition, /g/ is labelled as /f/ about three times more frequently than it is labelled as any voiced plosive). Such oddities are never explicitly discussed, as the data are only viewed through complex methods of data analysis that effectively smooth the results. **(2)** Their method of modulation used a noise that itself had a varying envelope, so that the resulting stimulus had envelope fluctuations that were a compound of the stimulus and the noise. In addition, the stimuli differed not only in time structure, but also in spectrum, because a pink noise was modulated instead of a white one (although it is not clear what importance this has for the listener). **(3)** Most importantly, the authors do not distinguish among the types of temporal information their stimuli contain (for example, they analyze the confusions from all three sets of stimuli with the same envelope features). From the point of view described above, the stimuli whose "envelopes" were extracted by low-pass filtering at 2 kHz had temporal information of all three kinds, while those filtered at 20 Hz had only envelope information. The 200 Hz filtered stimuli had envelope *and* periodicity information.

A pilot study has been done in an attempt to clarify these issues. A set of 12 VCVs (5 tokens per sound) were used. Speech "envelopes" were transformed into "signal-correlated noise" (Schroeder, 1968), which is equivalent to multiplying the envelopes by a *constant-amplitude* white noise, resulting in a signal whose instantaneous amplitude is identical

to that of the original signal, but whose spectrum is white[6]. Signal-correlated noise preserves the temporal information in the original signal, but eliminates its spectral characteristics. Four sets of stimuli were transformed into signal-correlated noise. Two of these (full-wave rectified speech low-pass filtered at 20 Hz and 2000 Hz, abbreviated as LO:2k, respectively) repeated conditions of Van Tasell et al. (although with a much greater bandwidth in the original speech signal, 20 kHz vs. Van Tasell et al.'s 4 kHz). A third used the unaltered speech signal (SP), while the fourth used half-wave rectified speech (HALF). Figures 1-3 show the waveforms and spectrograms for a test stimulus "ah-kah" in three conditions: natural speech, LO:20 and HALF. Figure 4 shows an expanded portion of the same stimulus starting from the aperiodic release burst and aspiration of the "k", and extending into part of the following periodic vowel for condition HALF. Note that although there is no spectral variation in the signal, much of the temporal complexity of the original speech is present in condition HALF.

Before going on to the empirical results, let us first consider the extent to which the each of these transformed signals contains information about the various phonetic features (summarized in Table 2). Three of the four signals retain information about all three phonetic features, therefore any differences in performance among these three must result from differences in the listeners' abilities to use the information in the form it is available. For condition LO:20, only for the manner feature should performance be similar to the other conditions - transmission of voicing should be considerably poorer and transmission of place nonexistent.

The broad pattern of the results was similar for two listeners who labelled each consonant in each condition 16-24 times. Figure 5 summarizes mean listener performance in terms of overall performance, and on the three main phonetic features. Overall performance was worst for LO:20 Hz and best for HALF. LO:2k and SP led to performances that were essentially the same in every respect, and in-between the other two conditions. The transmission of voicing information was relatively poor in LO:20 Hz, better in LO:2k and SP, and best in HALF. The same pattern was obtained for manner information, although the four conditions differed less (as expected from Table 2). Perhaps the most interesting result was the extent to which place of articulation was transmitted to the listeners, a feature that requires access to temporal fine-structure. For 3 of the 4 conditions, transmission of place information

6. There are two ways to think of the process of generating signal-correlated noise, both of which require a digitally-sampled signal. In one way, we take the digital signal and randomly flip (with a probability of 0.5) the polarity of each sample point. In the other way (alluded to already), the signal modulates a flat-spectrum noise created by generating a random series of plus and minus ones (or any arbitrary constant value). Clearly, the two techniques are identical. The latter "modulation" viewpoint is intuitively closer to the methods used in analogue hardware (and by Van Tasell et al.) while the former is closer to the way the process is actually programmed.

was very low (< 4%). Only in condition HALF was this exceeded with values of about 11%. Although still a weak cue, some information about place does seem to have been transmitted[7]. It may be that extended experience with such signals may lead to much improved performances. Furthermore, use of temporal fine-structure may be much more evident in tasks using stimuli with relatively long-lasting steady-state spectral characteristics (e.g., vowels).

These results support the view that temporal information may be divided into at least three aspects. Voicing information is weakest in LO:20 because performance must be based on the weak envelope cue of relative amplitude, and not the direct indication of periodicity/aperiodicity present in the other signals. As other studies (e.g., Faulkner et al., 1989) show that periodicity information with no envelope variations leads to a very good perception of voicing, it appears that periodicity information is more important than amplitude envelope for the perception of voicing.

Since condition SP is essentially what would be obtained if signal-correlated noise had been constructed from the full-wave rectified speech, the similarity of results between LO:2k and SP indicates that modulations in the signal above 2 kHz are not available to the listeners (consistent with the fact that 100% sinusoidally-amplitude-modulated white noise can only be distinguished from unmodulated noise up to frequencies of about 1-2 kHz, e.g., Bacon & Viemeister, 1985). Finally, the availability of temporal fine-structure found in HALF (and better perception of periodicity) may result because this signal has the lowest temporal density of the three signals which preserve fine-structure (see Figure 6). LO:2k and SP have roughly twice the density of modulations of the half-wave rectified signal. This makes temporal fine-structure in the more complex signals inaccessible to the listener due to insufficient temporal resolution. It may be that some further simplification would aid the listener in making more use of temporal fine-structure.

FINAL REMARKS

Clearly, the current study needs to be extended to further listeners, and to giving listeners more extended experience with these unusual signals. However, it is already apparent that the use of signal-correlated noise based on processed and unprocessed versions of speech sounds leads to two important advantages: (1) the ability to simulate the performance

7. Prideaux (1989) has done a more extensive study of this kind using 6 phonetically naive listeners and 6 listeners with explicit phonetic training (speech therapists, phoneticians and speech scientists). The pattern of results was similar for both groups, but the differences among conditions tended to be small, especially for the phonetically-naive group. It is likely that with more experience, listeners would have performed considerably better. Two listeners in the phonetically aware group (one of whom was one of the listeners in the pilot experiment) gave results essentially the same as those in Figure 5.

exhibited by users of single-channel implants in normal listeners, and (2) the ability to present signals to normal listeners that have only temporal information, in this way assessing the importance of temporal information in auditory processing. Signal-correlated noise is particularly useful in that it contains no amplitude fluctuations of its own, and so the fluctuations in the stimuli reflect only those in the original speech signals.

There are other ways that the experimental techniques could be improved. For example, low-pass filtering rectified speech to obtain stimuli with varying types of temporal information has some drawbacks. It does not cleanly distinguish temporal information of the three types (as their frequency regions overlap) nor is it possible to construct stimuli with only periodicity, and no envelope information. Two techniques could be used to avoid these limitations. Nonlinear smoothing methods (e.g., using the maximum value) with a time window locked to the period of voiced sounds should lead to more accurate extraction of envelope information. Secondly, it is a relatively simple matter to determine the presence or absence of voicing, extract fundamental frequency, and periods of frication from a digital signal (e.g., with the aid of a laryngograph, Fourcin, 1981) so as to construct stimuli that only indicate periodicity, without envelope or fine structure cues (as in Faulkner et al., 1989). Artificial manipulation of the relative amplitudes of different parts of signals (a type of "feature cross-splicing experiment") can also be used to clarify the dominance of envelope or periodicity cues.

In the end, a clearer understanding of temporal processes in audition will lead not only to a clearer understanding of normal hearing, but also to cochlear implant processing schemes which better exploit residual temporally-sensitive auditory abilities for speech reception. This applies as much to multi-channel systems, which have the possibility of temporal structure in each channel, as it does to single-channel systems which must rely totally on temporal information.

ACKNOWLEDGEMENTS

First thanks must go to Kerensa Prideaux who did some of the preliminary work for the pilot experiment as part of a student project. Mike Johnson, Andrew Faulkner, Bill Barry, Ginny Ball, Michael Ashby and Evelyn Abberton made useful comments in discussions, and/or suggested a number of changes to an earlier version of the manuscript. This work is supported by the Medical Research Council of the United Kingdom. Special travel grants from the Heinz and Anna Kroch Foundation, and the Royal Society, made attendance at the conference possible.

REFERENCES

Agelfors, E., & Risberg, A. (1987) *The identification of synthetic vowels by patients using a single-channel cochlear implant.* Proceedings of the

XIth International Congress of Phonetic Sciences, Tallinn 4:181-184. Also in (1987) Speech Transmission Laboratory -Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm 2-3:31-38.

Agelfors, E., & Risberg, A. (1989) *Speech feature perception by patients using a single-channel Vienna 3M extracochlear implant.* Proceedings of the Speech Research '89 International Conference, Budapest (Linguistics Institute of the Hungarian Academy of Sciences, Budapest) 149-152. Also in (1989) Speech Transmission Laboratory - Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm 1:145-149.

Bacon, S.P. & Viemeister, N.F. (1985) *Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners.* Audiology 24:117-134.

Bailey, P.J., & Summerfield, Q. (1980) *Information in speech: observations on the perception of [s]-stop clusters.* Journal of Experimental Psychology: Human Perception and Performance 6:536-563.

Blamey, P. J., Dowell, R. C., Brown, A. M., Clark, G. M., & Seligman, P. M. (1987) *Vowel and consonant recognition of cochlear implant patients using formant-estimating speech processors.* Journal of the Acoustical Society of America 82:48-57.

Burns, E.M., & Viemeister, N.F. (1976) *Nonspectral pitch.* Journal of the Acoustical Society of America 60/863-869.

Crystal, D. (1969) *Prosodic Systems and Intonation in English*, Cambridge: Cambridge University Press.

Dorman, M.F., Dankowski, K., McCandless, G. & Smith, L. (1989) *Identification of synthetic vowels by patients using the Symbion multi-channel cochlear implant.* Ear and Hearing 10:40-43.

Dorman, M.F., Raphael, L.J. & Isenberg, D. (1980) *Acoustic cues for a fricative-affricate contrast in word-final position.* Journal of Phonetics 8:39-405.

Dowell, R. C., Martin, L. F. A., Tong, Y. C., Clark, G. M., Seligman, P. M., & Patrick, J. F. (1982) *A 12-consonant confusion study on a multiple-channel cochlear implant patient.* Journal of Speech and Hearing Research, 25, 509-516.

Dowell, R. C., Seligman, P. M., Blamey, P. J., & Clark, G. M. (1987) *Evaluation of a two-formant speech- processing strategy for a multichannel cochlear prosthesis.* Annals of Otology, Rhinology and Laryngology, 96 (Suppl. 128), 132-134.

Dowell, R. C., Tong, Y. C., Blamey, P. J., & Clark, G. M. (1985) *Psychophysics of multiple-channel stimulation.* In R. A.. Schindler &

M. M. Merzenich (Eds.), Cochlear implants (pp. 283-290). New York: Raven Press.

Eddington, D.K. (1983) *Speech recognition in deaf subjects with multi-channel intracochlear electrodes*, in Cochlear Prostheses, an International Symposium, edited by C. W. Parkins & S. W. Anderson, Annals of the New York Academy of Sciences, 405.

Faulkner, A., Potter, C., Ball, G., & Rosen, S. (1989) *Audiovisual speech perception of intervocalic consonants with auditory voicing and voiced/voiceless speech pattern information.* Speech, Hearing and Language: Work in Progress, Phonetics & Linguistics, University College London, 3:85-106.

Fourcin, A. J. (1981) *Laryngographic assessment of phonatory function*, in Proceedings of the Conference on the Assessment of Vocal Pathology, edited by C. L. Ludlow & M. O'C. Hart, ASHA Reports 11 (ASHA, Rockville, Maryland).

Fry, D. B. (1968), *Prosodic phenomena*, in Manual of Phonetics, edited by B. Malmberg (Amsterdam: North Holland).

Fujimura, O. (1962) *Analysis of nasal consonants*, Journal of the Acoustical Society of America 34:1865-1875. Also in Readings in Acoustic Phonetics, (1967), edited by I. Lehiste (MIT Press, Cambridge, Mass.).

Gerstman, L. J. (1957) *Perceptual dimensions for the friction portions of certain speech sounds.* Unpublished doctoral dissertation. New York University.

Harris, K. S. (1958) *Cues for the discrimination of American English fricatives in spoken syllabes*, Language and Speech, 1, 1-7. Also in Acoustic Phonetics (1976), edited by D. B. Fry (Cambridge University Press, Cambridge).

Hazan, V. (1989) *Individual variability in the perception of cues to place and voicing contrasts in initial stops.* Speech, Hearing and Language: Work in Progress, Phonetics & Linguistics, Univesity College London, 3:123-142.

Hochmair-Desoyer, I. J., Hochmair, E. S., Fischer, R. E., & Burian, K. (1980) *Cochlear prostheses in use: Recent speech comprehension results.* Archives of Oto-Rhino-Laryngology, 229, 81-98.

Hochmair-Desoyer, I. J., Hochmair, E. S. & Stiglbrunner, H. K. (1985) *Psychoacoustic temporal processing and speech understanding in cochlear implant patients*, in Cochlear Implants, edited by R. A. Schindler & M. M. Merzenich (Raven Press, New York).

Howard, D. M. & Seligman, P. M. (1983) *Initial comparisons between two simple time-domain fundamental frequency detectors.* Speech, Hearing and Language: Work in Progress, Phonetics & Linguistics, University College London, 1:95-105.

Howell, P. & Rosen, S. (1983) *Production and perception of rise time in the voiceless affricate/fricative distinction.* Journal of the Acoustical Society of America 73:976-984.

Howell, P. & Rosen, S. (1987) *Perceptual integration of rise time and silence in affricate/fricative and pluck/bow continua.* In: The Psychophysics of Speech Perception, ed M.E.H. Shouten, pp. 173-180 (Dordrecht: Martinus Nijhoff).

Lehiste, I. (1970) *Suprasegmentals.* (MIT Press: Cambridge, Massachusetts).

Lisker, L, & Abramson, A. S. (1964) *A cross-language study of voicing in initial stops: Acoustical measurements.* Word 20:384-422.

Mermelstein, P. (1975) *Automatic segmentation of speech into syllabic units.* Journal of the Acoustical Society of America, 58, 880-883.

Merzenich, M. M. (1985) *UCSF cochlear implant device, in Cochlear Implants,* edited by R. A. Schindler & M. M. Merzenich (Raven Press, New York).

Miller, J. L. (1981) *Effects of speaking rate on segmental distinctions,* in Perspectives on the Study of Speech, edited by P. D. Eimas & J. L. Miller (Lawrence Erlbaum, Hillsdale, New Jersey).

Plomp, R. (1983) *The role of modulation in hearing,* in Hearing - Physiological Bases and Psychophysics, edited by R. Klinke and R. Hartmann (Springer-Verlag, Berlin).

Prideaux, K. (1989) *Temporal information in the identification of intervocalic consonants.* 4th year project in Speech Sciences, Department of Phonetics & Linguistics, University College London.

Repp, B. H., Liberman, A. M., Escardt, T. & Pesetsky, D. (1978) *Perceptual integration of acoustic cues for stop, fricative and affricate manner.* Journal of Experimental Psychology: Human Perception and Performance 4:621-637.

Rosen, S. & Ball, V. (1986) *Speech perception with the Vienna extra-cochlear single-channel implant: a comparison of two approaches to speech coding.* British Journal of Audiology 20:61-83.

Rosen, S., Walliker, J. R., Brimacombe, J. A. & Edgerton, B. E. (1989) *Prosodic and segmental aspects of speech perception with the House/3M*

single-channel implant. Journal of Speech and Hearing Research 32:93-111.

Sachs, M. B. & Miller, M. I. (1985) *Pitch coding in the auditory nerve: Possible mechanisms of pitch sensation with cochlear implants*, in Cochlear Implants, edited by R. A. Schindler & M. M. Merzenich (Raven Press, New York).

Sachs, M. B., Young, E. D. and Miller, M. I. (1983) *Speech encoding in the auditory nerve: Implications for cochlear implants*, in Cochlear Prostheses, an International Symposium, edited by C. W. Parkins & S. W. Anderson, Annals of the New York Academy of Sciences 405:94-114.

Seligman, P. (1987) *Speech-processing strategies and their implementation.* Annals of Otology, Rhinology and Laryngology, 96 (Suppl. 128), 71-74.

Schroeder, M. R. (1968) *Reference signal for signal quality studies.* Journal of the Acoustical Society of America 44:1735-1736.

Schinn, P. & Blumstein, S. E. (1984). *On the role of the amplitude envelope for the perception of [b] and [w]*, Journal of the Acoustical Society of America, 75, 1243-1252.

Soli, S. D. (1983) *The role of spectral cues in discrimination of voice onset time differences.* Journal of the Acoustical Society of America 73:2150-2165.

Stevens, K. N. (1980) *Acoustic correlates of some phonetic categories.* Journal of the Acoustical Society of America 68:836-842.

Stevens, K. N. (1981) *Constraints imposed by the auditory system on the properties used to classify speech sounds: Data from phonology, acoustics and psycho-acoustics.* In: The Cognitive Representation of Speech, ed. T. F. Myers, J. Laver & J. Anderson. Amsterdam: North Holland.

Stevens, K. N. & Blumstein S. E. (1981) *The search for invariant acoustic correlates of phonetic features*, in Perspectives on the Study of Speech, edited by P. D. Eimas & J. L. Miller (Lawrence Erlbaum, Hillsdale, New Jersey).

Stevens, K. N. & Klatt, D. H. (1974) *Role of formant transitions in the voiced-voiceless distinction for stops*, Journal of the Acoustical Society of America, 55, 653-659.

Summerfield, Q., Bailey, P. J., Seton, J. & Dorman, M. F. (1981) *Fricative envelope parameters and silent intervals in distinguishing "slit" and "split".* Phonetica 38:181-192.

Tyler, R. S., Tye-Murray, N., Preece, J. P., Gantz, B. J., & McCabe, B. F. (1987) *Vowel and consonant confusions among cochlear implant patients: Do different implants make a difference?* Annals of Otology, Rhinology and Laryngology, 96 (Suppl. 128), 141-144.

Umeda, N. (1975) *Vowel duration in American English*, Journal of the Acoustical Society of America, 58, 434-445.

Van Tasell, D. J., Soli, S. D., Kirby, V. M., & Widin, G. P. (1987) *Speech waveform envelope cues for consonant recognition.* Journal of the Acoustical Society of America, 82, 1152-1161.

White, M. W. (1983) *Formant frequency discrimination and recognition in subjects implanted with intracochlear stimulating electrodes*, in Cochlear Prostheses, an International Symposium, edited by C. W. Parkins & S. W. Anderson, Annals of the New York Academy of Sciences 405:348-359.

Table 1. The role of various temporal features of speech in linguistic contrasts. The number of "*"s indicate the extent to which a particular feature operates in a particular linguistic contrast. Three "*"s indicate that the temporal feature conveys strong cues to the contrast, whereas a blank space indicates very weak or nonexistent cues.

			TEMPORAL FEATURE		
			envelope	periodicity	fine-structure
LINGUISTIC CONTRAST	segmental	manner	***	**	*
		voicing	*	***	**
		place			***
		vowel quality	*		***
	prosodic	tempo, rhythm	***		
		syllabicity	***		
		stress	*	***	
		intonation		***	

Table 2. Information about the three main phonetic features relating to consonants (manner, voicing and place) contained (at least theoretically) in the temporal structure of the four transformed speech signals. As before, the number of "*"s indicate the extent to which a particular signal contains information about a particular phonetic feature.

			TEMPORAL FEATURE(S) IN EACH SIGNAL			
			envelope	envelope, periodicity & fine-structure		
segmental	contrast		LO:20	LO:2k	SP	HALF
		manner	***	***	***	***
		voicing	*	***	***	***
		place		***	***	***

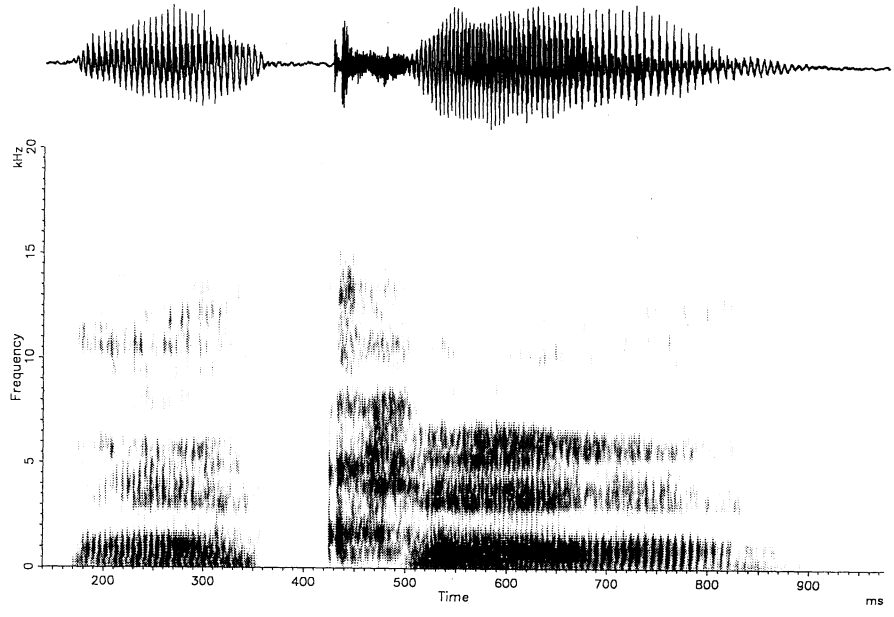

Fig. 1 : Speech pressure waveform and spectogram of the utterance "ah-kah". The signal was pre-emphasized to show higher frequencies better. For the spectrogram (as well as those that follow), and analyzing bandwidth of 400 Hz was used, with a display range of 30 dB. Note the vertical striations in the spectrogram that indicate vocal fold vibration during the vocalic intervals, and the lack of them during the voiceless consonantal gesture.

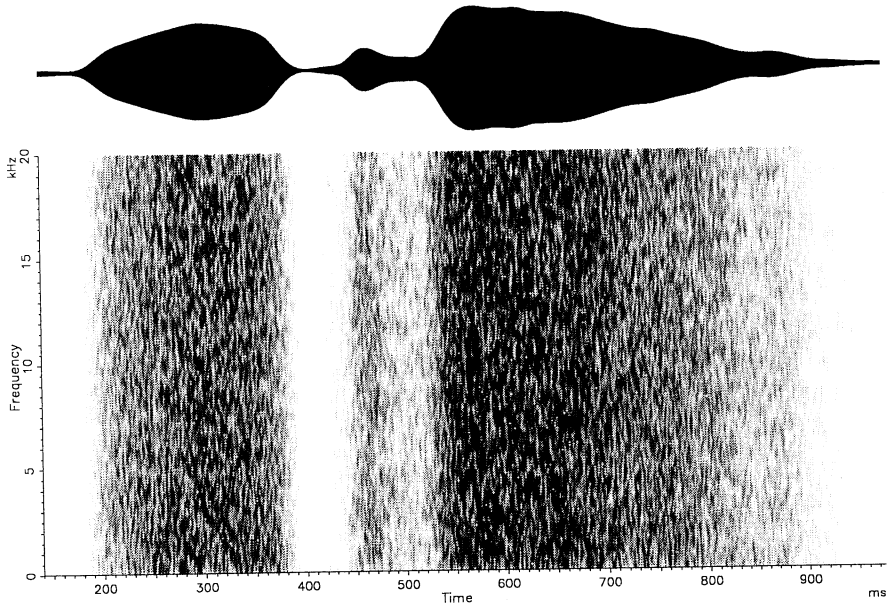

Fig. 2 : Waveform and spectrogram of the utterance "ah-kah" in condition L0:20. Note the lack of spectral structure in the spectrogram.

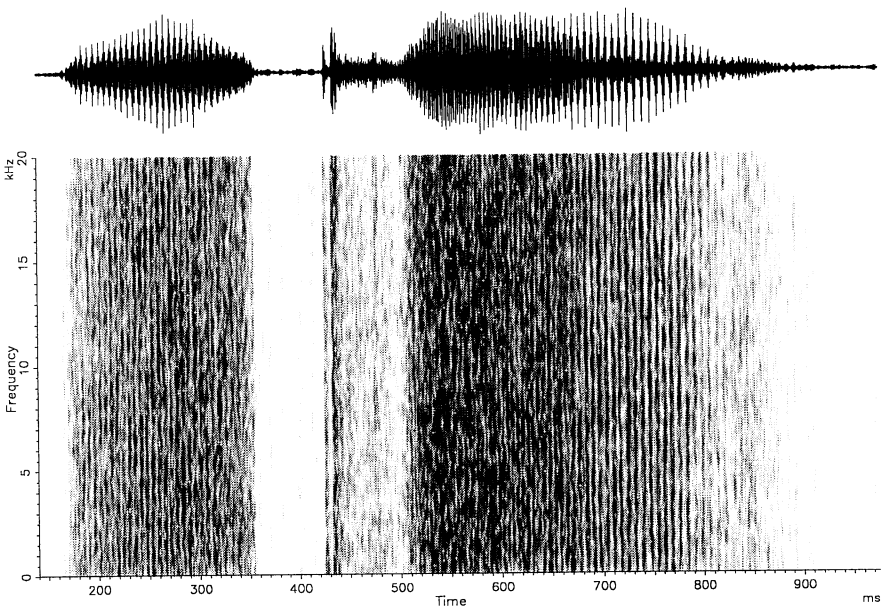

Fig. 3 : Waveform and spectrogram of the utterance "ah-kah" in condition HALF. Note the lack of spectral structure, along with the presence of much temporal structure. In particular, the presence of voicing is again indicated by vertical striations in the spectrogram.

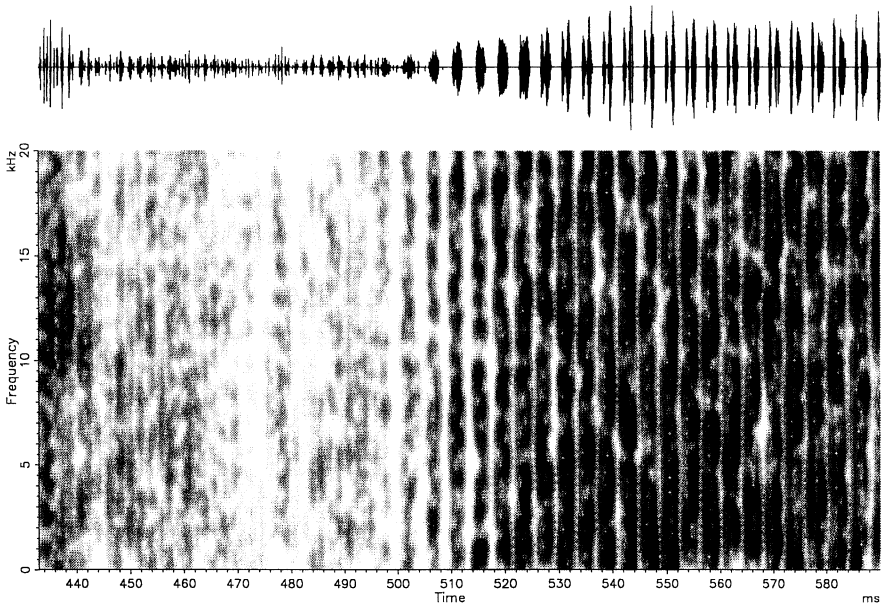

Fig. 4 : Waveform and spectrogram of the release burst and aspiration and part of the following vowel of the utterance "ah-kah" in condition HALF. Note the lack of spectral structure, along with the presence of much temporal structure.

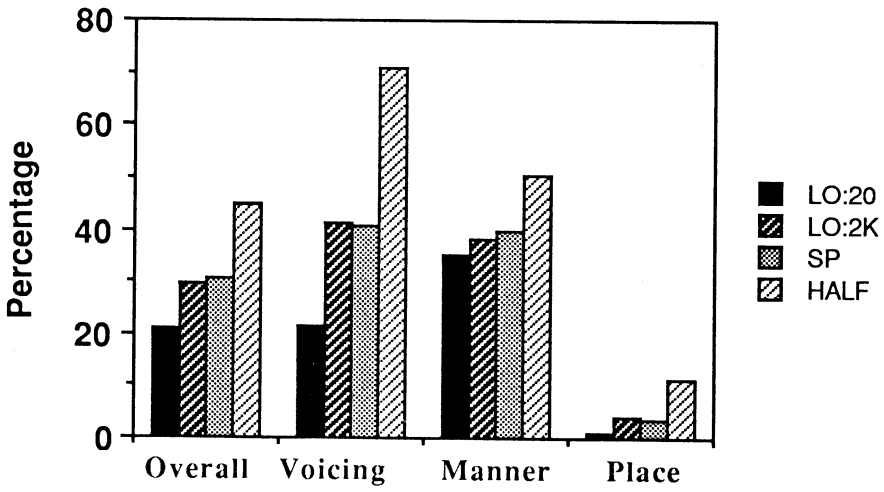

Fig. 5 : Mean results from two normal listeners requiring the identification of intervocalic consonants in which temporal, but not spectral structure is present. These values were determined by first summing confusion matrices over each listener, calculating the appropriate statistics, and then taking a simple mean. One listener identified each stimulus 16 times in each condition, while the other identified each 24 times, making a total of 40 observations per stimulus per condition.

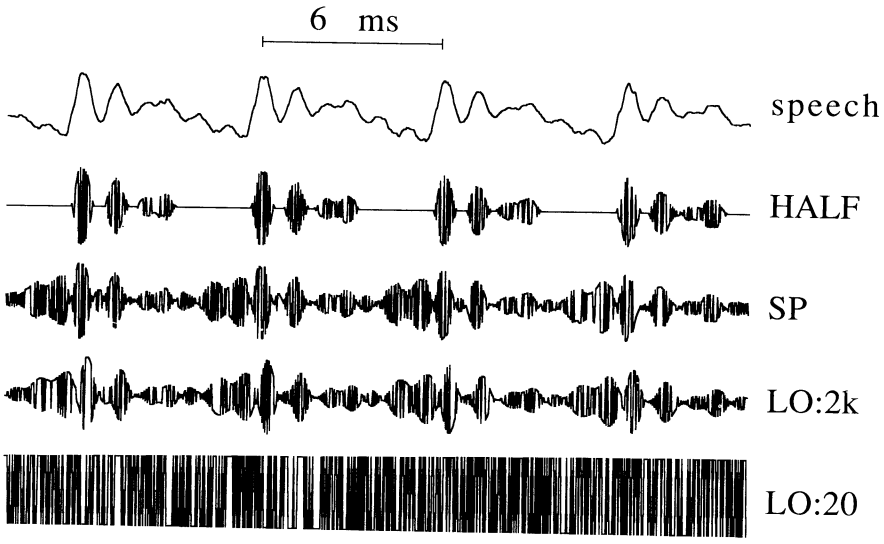

Fig. 6 : A short segment of the first vowel in "ah-kah" showing the original speech-pressure waveform (top) as well as the waveforms obtained in conditions HALF, SP, LO:2k and LO:20. Of the three transformed signals with temporal complexity related to the vowel periodicity and waveshape, HALF is less temporally dense than the other two signals. LO:20 displays none of the periodicity and fine-structure of the original speech.