Vowel-onset detection by vowel-strength measurement, cochlear-nucleus simulation, and multilayer perceptrons

Vowel-onset detection by vowel-strength measurement, cochlear-nucleus simulation, and multilayer perceptrons

Reinier W. L. Kortekaas and Dik J. Hermes
Institute for Perception Research/IPO, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Georg F. Meyer
Department of Computer Science, University of Keele, Keele, Staffordshire ST5 5BG, United Kingdom

(Received 1 November 1994; revised 18 July 1995; accepted 24 July 1995)

An algorithm for detection of vowel onsets in fluent speech was presented by Hermes [J. Acoust. Soc. Am. **87**, 866–873 (1990)]. Performance tests showed that detection was good for fluent speech, although the parameter settings had to be modified for application to well-articulated speech. One of the purposes of the algorithm was application to speech by deaf persons, for which it failed completely. In order to improve the algorithm and to make it more generally applicable, two alternative detection strategies have been explored in the present study. These strategies were (a) simulation of transient-chopper responses in the cochlear nucleus and (b) training of multilayer perceptrons. Two large databases of read speech have been used for performance comparison of the original algorithm and the new strategies. The strategy based on simulating cochlear-nucleus responses is found both to result in a higher false-alarm rate than the original algorithm and to be rather level dependent. On the other hand, the performance of a multilayer-perceptron network, trained on mel-scaled spectra, is comparable to the performance of the Hermes algorithm. In more general terms, the results suggest that temporal information on intensity and (rough) spectral envelope are important for human vowel-onset detection behavior. Information on harmonicity can be used as a secondary source of information to avoid detection of mainly unvoiced, nonvowel onsets.

PACS numbers: 43.72.Lc, 43.70.Fq, 43.64.Bt, 43.71.An

INTRODUCTION

A prominent characteristic of speech signals is the presence of simultaneous frequency modulation (FM) and amplitude modulation (AM). On a suprasegmental level, FM can be found in the pitch contour, whereas AM is present in the syllabic structure. The corresponding modulation frequencies, disregarding phenomena like microintonation, are typically low: In general, the upper bounds are 8 Hz for FM and 5 Hz for AM (e.g., Plomp, 1984). On the segmental level, FM and AM are most prominent in the so-called ''fast transitions'' that occur in the succession of two phonemes or in diphthongs. The rate of those modulations often exceeds by far the rate of suprasegmental FM and AM. There is growing evidence that fast transitions are important for phoneme recognition. Especially in the case of a plosive-vowel combination, a short portion of the speech signal (typically 20–40 ms) appears to contain sufficient information for determination of the place of articulation of the consonant (e.g., Kewley-Port *et al.*, 1983) or the identity of the vowel (e.g., Tekieli and Cullinan, 1979). In general, much perceptually relevant information is present in speech portions which show substantial spectral change within a short time interval (Strange *et al.*, 1983; Furui, 1986; Nossair and Zahorian, 1991).

Hermes (1990) defines vowel onsets perceptually as the moment in the syllable at which a listener starts to perceive the vowel. He determined vowel onsets by using a so-called gating paradigm ('t Hart and Cohen, 1964) in which short, windowed portions of the speech signal are isolated and listened to. Using this paradigm, a phonetician can determine vowel onsets with an average accuracy of at least 20 ms. Agreement among different phoneticians will generally be high, but may vary for different phonemic contexts. In the Hermes (1990) study, vowel onsets are assumed to coincide with speech portions which show large increments in intensity in separated frequency channels of the auditory system. As such, vowel onsets can be conceived of as a subset of the class of fast transitions.

House (1990) showed that speech segments with considerable spectral change play an important role in tonal perception in speech. The perceptual identity of a pitch movement, i.e., which syllable it accentuates and how much it contributes to the prominence of the syllable, depends on the temporal position of the pitch movement with respect to those segments with considerable spectral change. Of these segments, the vowel onset appears to play the most important role ('t Hart and Cohen, 1973; 't Hart and Collier, 1975; House, 1990). Furthermore, vowel onsets appear to play a primary role in perceiving rhythm in speech (Rapp, 1971; Allen, 1972; Cole and Scott, 1973; Eriksson, 1991). In phonology, the vowel onset corresponds to the transition from syllable onset to syllable rhyme. At the syllable level, this division is generally believed to be basic to the structure of the syllable (Pike, 1947; Selkirk, 1982). The theoretical issue of the importance of the vowel onset for speech perception will be readdressed in Sec. III.

In spite of these arguments concerning the relevance of vowel onsets to speech perception, little attention has been paid to the automatic detection of vowel onsets. Hermes

(1990) introduces the concept of ''vowel strength,'' which is a measure for the presence of both a formant structure, i.e., with pronounced maxima, as well as a harmonic structure, i.e., a line structure, in the amplitude spectrum. Strong increments in vowel strength are supposed to signal the presence of a vowel onset. Based on these ideas, Hermes (1990) presents an algorithm for the automatic detection of vowel onsets in natural speech, which we will refer to here as vowel-strength measurement (VSM). The algorithm was found to perform quite satisfactorily for fluent speech: Approximately 10% of vowel onsets were missed, most of which before unaccented schwas. For well-articulated isolated words, some parameter settings of the algorithm had to be changed for a satisfactory performance; otherwise, the number of false alarms increased unacceptably. The algorithm was applied in a teaching system of intonation to profoundly deaf persons (Kaufholz, 1992). In this case as well, the performance of the algorithm was judged to be unacceptable. As the aim of this system was to improve the intonation of deaf persons, the system could not be optimized for deaf speech. Two independent strategies for improving VSM were attempted: First, psychophysically inspired processing stages were introduced (te Rietmole, 1991), and second, cepstral analysis was included (Kaufholz, 1992). Neither strategy could substantially reduce the number of missed onsets without simultaneously increasing the number of false alarms. More generally, the relative contributions of physical speech-signal characteristics like intensity, harmonicity, and spectral content, remained unclear.

This paper will describe further investigations of automatic vowel-onset detection by comparing the performance of VSM to two alternative automatic detection strategies. The aim of the investigations is both to obtain more insight into the underlying signal dimensions which play a role in human vowel-onset detection, and to develop a better detection scheme. The rationale for comparing different schemes is that such a method may help to derive what information is required to model human vowel-onset detection. The paper will focus on the description and evaluation of the alternative strategies. In the evaluation, both missed-onset and false-alarm rates as well as phonemic context of occurrence will be analyzed.

As alternative vowel-onset detection strategies we have chosen (a) to apply a model of the cochlear nucleus (CN) developed by Meyer (1993a, b) and (b) to train multilayer perceptron (MLP) networks. In the case of the CN model, a detection scheme will be presented that is based on simulated transient-chopper responses. This detection scheme will be referred to as cochlear-nucleus simulation (CNS). The MLP networks have been trained both with the simulated transient-chopper responses and with conventional mel-scaled amplitude spectra. We have compared VSM, CNS, and the MLP networks by means of performance tests using two speech databases. Both databases consisted of Dutch read speech uttered by nonprofessional speakers: 18 and 24 speakers, respectively, with an equal number of male and female speakers. Recording conditions of the databases were good and excellent, respectively. The results give support to the hypothesis that the important signal dimensions for human vowel-onset detection are both the increment of intensity as well as the spectral envelope, and that harmonicity serves as an additional source of information. The CNS detection scheme yields missed-onset rates slightly higher than VSM, and false-alarm rates that are considerably higher. Moreover, the performance of the CNS scheme is seen to be substantially level dependent. On the other hand, the performance of a detection scheme based on an MLP network, with mel-scaled spectra as input, is competitive with the performance of VSM.

I. MODELS

In this section, the detection schemes will be presented: first, the VSM algorithm will be briefly reviewed. Second, both the auditory model and the corresponding vowel-onset detection scheme CNS will be discussed, and finally, the MLP-network architecture will be presented.

A. Vowel-strength measurement

The VSM algorithm is based on measurement of vowel strength. This measure expresses (a) the weighted contribution of the harmonics to the pitch of a speech segment,[1] and (b) the degree to which a formant pattern is present in the preprocessed amplitude spectrum of a pitch period within that segment. Vowel strength is measured every 10 ms. Both (a) and (b) show a high correlation to intensity so that vowel-onset detection in VSM is based on spectral envelope, harmonicity, and intensity.

In VSM, vowel onsets are associated with rapid increments in vowel strength. Such increments are detected by finding the maxima of the smoothed derivative of the sequence of vowel-strength measurements. The impulse response of the smoothed-derivative filter has a bipolar character; it is the sum of two Gaussians shifted in time and of opposite sign, starting with a positive excursion. The effective duration of the filter is approximately 100 ms. A similar filter is applied in the CNS detection scheme and the MLP networks. An example of the time course of vowel strength is given in Fig. 1 (PM-sentence 13 ''Eindelijk kwam de trein op gang,'' see Sec. II A). The smoothed-derivative filter is shown in the inset of Fig. 1. For further details of the VSM algorithm, the reader is referred to Hermes (1990).

B. Detection based on simulated transient-chopper responses

If vowel onsets play a role in human speech perception, it is reasonable to expect that the information required for this process is encoded in the auditory pathway. The auditory model (Meyer, 1993a, b; Ainsworth and Meyer, 1994) applied in this study comprises simulations of responses in the first two stages in the neural auditory pathway: the cochlear or auditory nerve (AN) and the anteroventral cochlear nucleus (AVCN). For these stages, however, no physiological observations have been reported in the literature that demonstrate any enhancement of vowel onsets in neural responses (Kortekaas and Meyer, 1994). Nevertheless, simulation of neural encoding of (speech) signals may be appropri-

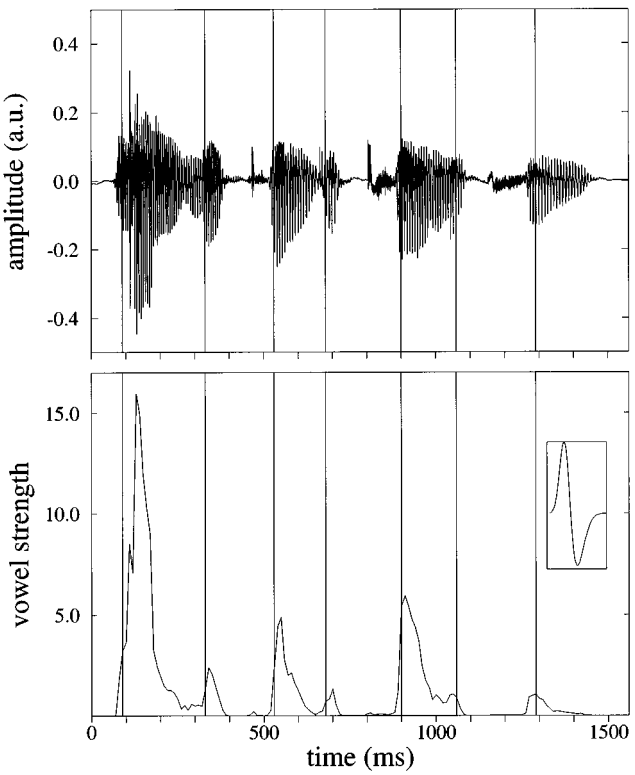

FIG. 1. Waveform of PM-sentence 13 (top) and corresponding vowel strength, as measured by VSM, as a function of time (bottom). Aurally detected vowel onsets are marked by vertical lines. The impulse response of the smoothed-derivative onset filter is shown in the inset of the bottom panel. Note that the time axes of vowel strength and the onset filter do not match; the effective duration of the onset filter is approximately 100 ms.

ate for automatic vowel-onset detection; the spectrotemporal resolution of such simulations is approximately the same as that of the auditory system. This means that such simulations may contain information which is perceptually more important. Moreover, several authors have demonstrated the perceptual importance of onsets and offsets, which are enhanced in simulated neural encoding and therefore more easily detectable (e.g., Darwin, 1984; Summerfield and Culling, 1992). Furthermore, spectral contrast is enhanced, too, which will be important for tracing resolved harmonics and formant frequencies.

The initial motivation for investigating simulations of peripheral auditory processes was based on the observation of overshoot phenomena in the AN, like short-term adaptation (e.g., Smith and Zwislocki, 1975; Eggermont, 1985). The hypothesis was that, due to short-term adaptation, rapid intensity increments and decrements in auditory channels are strongly enhanced in the firing profiles of cochlear nerve fibers. In the present context, rapid increments correspond to vowel onsets. The idea was that vowel onsets can be detected by measuring such strong simultaneous increments in channels corresponding to formant regions. It should be noted that an alternative approach would be to concentrate on phase locking in the auditory nerve. Several authors have demonstrated that the spectral content of speech sounds is preserved in the phase-locked activity of cochlear-nerve fibers (e.g., Delgutte and Kiang, 1984a, b, c; Carney and Geisler, 1986).

Nerve fibers generally show a limited dynamic range, normally extending 20–30 dB of differential sensitivity (e.g., Evans and Palmer, 1980). Steady-state stimulation at levels that exceed this range results in saturation of the fiber and as a consequence, loss of differential coding. The dynamic ranges of single nerve fibers generally are too small to provide differential coding over the whole dynamic range of speech. On the other hand, the auditory system seems to be provided with a continuum of nerve fibers with different thresholds and dynamic ranges (Liberman, 1978). Often a categorization of nerve fibers into a low- and a high-threshold population is made. A combination of the two populations seems necessary for coding spectral information in terms of discharge rate over the whole dynamic range of speech.

Such a combination is found for the so-called stellate cells in the AVCN (e.g., Rhode and Greenberg, 1992). Stellate cells receive excitatory input from both low- and high-threshold AN fibers whose characteristic frequencies are within 1 Bark of the characteristic frequency of the stellate cell (Rhode and Smith, 1986; Blackburn and Sachs, 1990; Rhode and Greenberg, 1992). Moreover, the cell receives inhibitory input from a relatively wide receptive field. The response patterns that were physiologically recorded from stellate cells are often characterized as ''transient chopper'' or ''chop-T.'' Such a response pattern is characterized by initial regularity of discharge (''chopping'') followed by a rapid transition to irregularity. Under stimulation with pure tones, the dynamic range of chop-T cells is comparable to the limited dynamic range of AN fibers. Nevertheless, physiological recordings have shown that discharge profiles of chop-T cells represent the spectrum of (speech) sounds over a wide dynamic range. Therefore, simulation of chop-T responses seems to be appropriate for vowel-onset detection (Kortekaas *et al.*, 1994). Blackburn *et al.* (1990) showed, for instance, that the formants of the synthetic vowel /ɛ/ are represented with high contrast in chop-T discharge rates over a dynamic range of 35- to 75-dB sound-pressure level. Such contrasts were not observed in the rate profiles of AN fibers. These differences in spectral contrast can be modeled by lateral inhibition which is present in the case of chop-T neurons, but not in the case of AN fibers.

1. Auditory model

This section briefly describes the auditory model which consists of a peripheral part and a chop-T-response simulation part (Meyer, 1993a, b). The peripheral part consists of the following stages:

(1) Gamma-tone filterbank with 32 4th-order IIR filters (De Boer, 1969; Darling, 1991). Center frequencies range from 0.1 to 4.6 kHz at 0.5 Bark spacing. Each channel has a bandwidth of one equivalent rectangular bandwidth (ERB) (van Compernolle, 1991).
(2) Filter-output scaling for:
Human hearing-threshold adjustment (Fay, 1988).
Dynamic-range extension (see below).
(3) Inner-hair-cell model (Meddis, 1986, 1988).

(4) Spike generation on the basis of expected firing rates.

The hearing thresholds are calculated for each channel by a polynomial fit to the data reported in Fay (1988).[2] The dynamic-range-extension stage is introduced for simulation of two populations of nerve fibers; instead of adjusting the parameters of the Meddis (1988) model, the signals that drive the inner-hair-cell model are scaled. The two populations are specified as follows, where relative firing threshold denotes the absolute firing threshold of the fiber relative to the absolute hearing threshold (i.e., sensation level):

Low-threshold fibers: relative fiber threshold of 0 dB, dynamic range of 30 dB, and spontaneous activity of 50 spikes s^{-1};
High-threshold fibers: relative fiber threshold of 15 dB, dynamic range of 50 dB, and spontaneous activity of 15 spikes s^{-1}.

Simulation of CN responses is based on a point-neuron model. The membrane potential is controlled by the Goldman–Hodgkin–Katz equation (e.g., Brown, 1991) as a function of concentration gradients and membrane permeability. Each simulated neuron receives excitatory input from neurons having the same center frequency. In addition, the neuron receives inhibitory input from neurons which have center frequencies between 1 Bark below, and 2 Bark above the neuron's center frequency. Both for excitation and inhibition, afferent neurons are from both populations. Instead of generating action potentials for the chop-T simulation, we concentrate on the extracellular potential of the neuron relative to its firing threshold. We will refer to this potential as ''activity.'' The output of the chop-T simulation is the activity, as a function of time, of an array of 23 neurons. Best frequencies of those neurons are in the range from 0.2 to 2.6 kHz, with 0.5-Bark spacing. For more details on the model, the reader is referred to Meyer (1993a, b) and Ainsworth and Meyer (1994).

2. Vowel-onset-detection scheme

The vowel-onset-detection scheme was developed with the aim of applying phonetic knowledge to the process of detecting vowel onsets in simulated chop-T responses. We will refer to this scheme as cochlear nucleus simulation (CNS) in the following. In the scheme, the simulated neuron activity is averaged over two frequency bands which roughly correspond to the regions of the first and second formant (see Fig. 2). The corresponding best frequencies are 0.2–1.1 kHz for the first formant, and 0.9-2.6 kHz for the second-formant band, respectively. These two bands can be conceived as a rough representation of the spectral envelope within the first and second formant region. The underlying idea of defining two formant areas is the assumption that vowel onsets are characterized by simultaneous and strong increase of activity in these two bands. The scheme traces such increases of activity (''vowel-onset candidates'') and applies a number of criteria to discard onsets other than vowel onsets.

The averaged activity is low-pass filtered by means of leaky integration to obtain the signals $A_L(t)$ and $A_H(t)$ for the first- and second-formant band, respectively. The -3-dB point of the LP filter is at $\sim$25 Hz. Subsequently, we take the

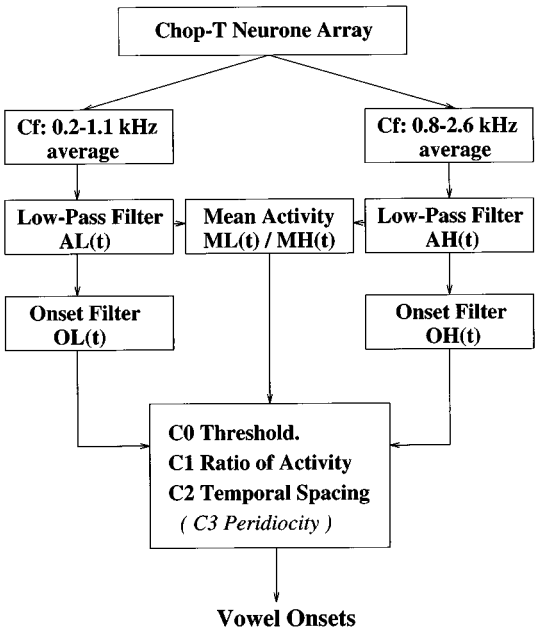

FIG. 2. Schematic representation of the CNS (and CNS–ACF) detection scheme.

smoothed derivative of $A_L(t)$ and $A_H(t)$, as described in Sec. I A, which results in the signals $O_L(t)$ and $O_H(t)$. Vowel-onset candidates are found at those instances at which (a) both $O_L(t)$ and $O_H(t)$ are greater than zero, (b) $O_L(t)$ has a local maximum. The condition that $O_H(t)$ should also have a local maximum was found to be too strict.

A vowel-onset candidate detected at time t_c should match the following criteria (parameter settings will be given below).

a. C0 threshold. To exclude irrelevant fluctuations of activity in $A_L(t)$, the criterion reads:

$$O_L(t_c) \geqslant O_{TH},$$

where O_{TH} is a parameter. A similar criterion is applied in VSM.

b. C1 ratio of activity. In order to discard candidates that signal phoneme classes other than vowels, the following criterion is introduced:

$$\alpha \leqslant \frac{M_L(t_c)}{M_H(t_c)},$$

where α is a parameter and $M_L(t_c)$ and $M_H(t_c)$ denote the mean activity in $A_L(t)$ and $A_H(t)$ over 45 ms following the onset at t_c. The lower bound α is introduced to discard onsets of fricatives which have concentration of spectral energy in the high region. An upper bound for the ratio of activity, which could exclude onsets for nasals, was found not to contribute significantly.

c. C2 temporal spacing. Some phoneme classes, e.g., liquids and semivowels, yield multiple candidates during an interval of continuous increase of activity, i.e., $O_L(t) > 0$ for all t within the interval. This criterion consists of two parts:

(i) Within such an interval, a vowel onset detected at t_1 is discarded if a subsequent vowel onset is detected at t_2, with $t_2 > t_1$.

(ii) If a candidate at t_2 is separated from a preceding vowel onset at t_1 by a period of decrease or absence of activity, then the candidate at t_2 is accepted if

$$t_2 - t_1 > \Delta,$$

where t_1, t_2 like above and Δ is a parameter. The vowel onset at t_1 is not discarded. A similar criterion is applied in VSM.

Unlike VSM, the CNS scheme does not include information about the harmonicity of the signal. Modulation transfer functions measured for chop-T responses show that phase locking to the AM envelope is preserved up to approximately 400 Hz (e.g., Rhode and Greenberg, 1992). This means that information about harmonicity of the speech signal can be derived from a broad frequency range. A rather *ad-hoc* solution to incorporate a harmonicity criterion into CNS is to calculate the short-term autocorrelation of the activity of each chop-T neuron. The autocorrelation is calculated for a 10-ms signal window. The individual autocorrelations are combined to obtain the summary autocorrelogram of the whole neuron array (e.g., Meddis and Hewitt, 1991). The magnitude of the maximum of the summary autocorrelogram is taken to represent the ''strength of harmonicity.'' This measure generally shows a high correlation to the instantaneous level of the input signal. In this extended scheme, referred to as CNS–auto-correlogram function (CNS–ACF) in the following, a criterion is introduced for periodicity:

d. C3 periodicity. For each vowel-onset candidate, the corresponding strength of periodicity should be above ACF_{TH}% of the maximum strength of periodicity observed in the utterance. Here, ACF_{TH} is a parameter.

An example of the signals $A_L(t)$, $A_H(t)$, and the autocorrelation peak is given in Fig. 3 for PM-sentence 13 (''Eindelijk kwam de trein op gang;'' see Sec. II A). Note that the C3 criterion requires that the whole utterance is analyzed before vowel onsets can be detected. Instead, the maximum strength of periodicity could also be determined for a constant interval of, e.g., 200 ms. Such an interval may require, however, that (a) pauses in utterances can be detected and (b) that the signal-to-noise ratio (SNR) is always high.

In summary, vowel-onset detection in the CNS scheme is based on changes of intensity and (rough) spectral envelope. In addition to these characteristics, detection in the CNS–ACF scheme is based on estimation of harmonicity.

3. Scaling of input signals

The computational model of the auditory periphery and the cochlear-nucleus responses is nonlinear, which makes input scaling necessary. Optimizing and testing the CNS and CNS–ACF schemes (and MLP networks) was done at 35, 55, and 75 dB SPL, where the root-mean-square of each of the sentences was normalized. Note that both schemes are partially based on intensity-level information, but that the rate-intensity functions of the simulated neurons behave nonlinearly.

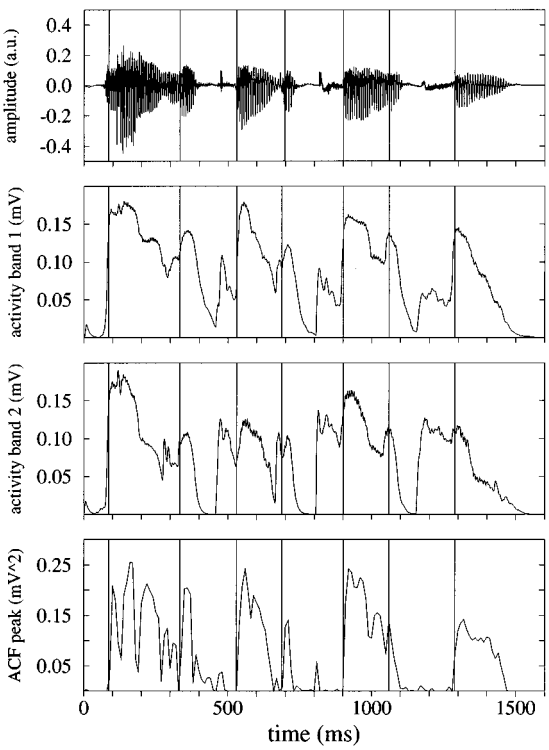

FIG. 3. Waveform of PM-sentence 13 (top), activity in band 1 and band 2 at 55 dB SPL (middle panels), and magnitude of the peak of the summary autocorrelogram of the chop-T neuron array at 55 dB SPL (bottom). Aurally detected vowel onsets are marked by vertical lines.

4. Parameter setting

The parameters presented above were optimized for the T-sentence database containing 377 vowel onsets (see Sec. II A). Setting all parameters to zero, i.e., accepting all vowel-onset candidates, resulted in 356 correct detections and 136 false alarms (at 55 dB SPL input level; see Sec. I B 3). Using the criteria mentioned above, the vowel-onset-detection scheme has to perform a ''yes–no'' task for each vowel-onset candidate. A method for analyzing the performance of the scheme as a function of parameter settings is the receiver–operating-characteristic (ROC) curve (e.g., Green and Swets, 1966). By setting all other parameters to zero, individual parameters were optimized by finding the value that optimally reduces the number of false alarms, while keeping the number of correct detections almost unaffected. This means that the combination of the different optimized criteria may drastically reduce the false-alarm rate, provided the different criteria affect different vowel-onset candidates. The parameter settings listed in Table I were derived by finding optimal values compromising for both 55 and 75 dB SPL input level.

Figure 4 shows ROC curves for CNS and CNS–ACF where each of the parameters is varied while all other param-

TABLE I. The parameters of the CNS and CNS–ACF schemes.

O_{TH}	0.00075 mV/ms
α	0.8
Δ	45 ms
$\mathrm{ACF}_{\mathrm{TH}}$	10 %

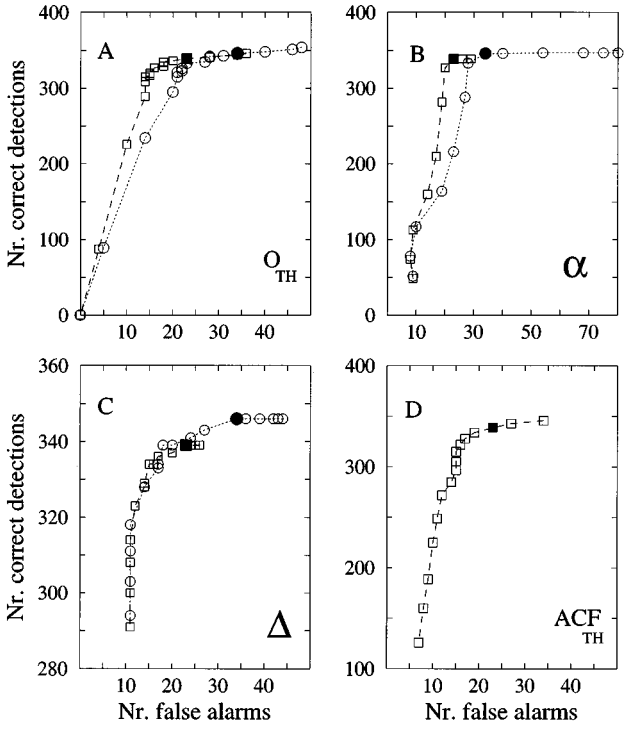

FIG. 4. ROC curves for each of the parameters of CNS (○) and CNS−ACF (□). Filled symbols represent the parameter settings used in the evaluations and listed in Table I. In all cases, the smallest parameter values correspond to the rightmost data points on the ROC curve. (a) ROC curve for parameter O_{TH} with parameter range of 0 to 0.003 mV/s in steps of 0.00025 mV/s. Additional values are 0.004, 0.008, 0.016, and 0.032 mV/s. (b) ROC curve for α where the range is 0 to 1.5 with a stepsize of 0.05. (c) ROC curve for Δ where values are set in the range 5 to 100 ms with increments of 5 ms. (d) ROC curve for ACF_{TH} plotted for the range 0% to 75% in steps of 5%. These data are obtained for the T-sentence database at 55 dB SPL (see Sec. II A).

eters are set to the optimal values as listed in Table I. These curves show the sensitivity of both schemes to variation of each of the parameters. This sensitivity can be derived to a first approximation from the area under the ROC curve (Green and Swets, 1966). It should be noted that the ROC curves are plotted with raw numbers as units, whereas ROC curves usually display probabilities. Moreover, for all parameters except O_{TH}, only part of the ROC curve is shown. In general, the individual optimal parameter settings determined by the ROC-curve method described above are seen to be also appropriate in case all criteria are applied. This indicates that the criteria can be conceived of as being more or less independent. From Fig. 4(b) it can be derived that the contribution of the C3 criterion is important in case no spectral-envelope information, i.e., the C1 criterion, is used. If the C1 criterion is applied then the C3 criterion contributes by additionally discarding some false alarms while keeping the correct-detection rate almost unaffected [see Fig. 4(d)].

C. Training multilayer perceptron (MLP) networks

MLP networks have proven to be robust techniques for pattern classification in speech-recognition tasks (e.g., McCulloch and Ainsworth, 1988; Markowitz, 1993). A weak point of the CNS(−ACF) detection scheme is that the decision boundary for the vowel versus nonvowel categories may not be optimal. In other words, the criteria defined on the basis of general phonetic knowledge may not be optimal for separating the two categories. In this respect, MLP networks are used (a) to determine whether the information required for vowel-onset detection is present in the chop-T representation, and (b) to investigate whether this information can be retrieved from conventional speech-analysis techniques. MLP networks are used in this study as vowel versus nonvowel classifiers, where two classification experiments have been performed:

(i) MLP networks have been trained with the chop-T representation for the following input-pattern configurations:
(a) the full-resolution, 23-channel representation;
(b) the two-formant-band representation as applied in CNS(−ACF);
(c) a single unit being the sum of all 23 channels.
In the case of (b), performance results inform about the amount of information required for vowel-onset detection with respect to spectral envelope. In the case of (c), only intensity information was supplied to the network.

(ii) MLP networks have been trained with conventional mel-scaled amplitude spectra. The choice of mel-scaled spectra was motivated by the similarity of these spectra to the simulated cochlear-nucleus representation except for the nonlinearities of the auditory model. The mel-scaled spectra consisted of 23 points covering approximately the same frequency range as in VSM and CNS(−ACF), namely 200–2600 Hz. Three input-pattern configurations have been investigated:
(a) the rms of each utterance was normalized so that individual spectra contained intensity-level information;
(b) each mel-scaled spectrum was normalized so that no intensity-level information was used;
(c) all components of individual mel-scaled spectra were summed to obtain a measure of the instantaneous level.
For input patterns under condition (a), all spectra contained intensity-level information with respect to the whole utterance. In contrast, input patterns under (b) did not contain intensity-level information but only contained spectral envelope information. Finally under condition (c), where the rms of the utterance was normalized, patterns only consisted of a measure of the instantaneous level.

1. Network architecture

In all classification experiments, the output of the MLP networks consisted of a single unit representing the presence of a vowel. The unit's output range was between zero and one, where zero indicates that the presence of a vowel, given the input pattern, is highly unlikely. In the case of both the mel-scaled spectra and the chop-T representation, input patterns consisted of 1, i.e., the sum of spectral components, or 23 units, i.e., the full resolution. Performance was evaluated for the number of hidden units ranging between 0 and 10. If no hidden units were used, the network did not learn, whereas ten hidden units caused "overtraining." Best results on the training database (see Sec. II A) were obtained for two to five hidden units: Except for the single input condition,

where the network incorporated two hidden units, all results to be described below were obtained with networks having five hidden units.

2. Network training

All input patterns were calculated over 25.6-ms-long time slices by either calculating an FFT of the windowed speech signal, or by binning the activity within channels in the chop-T representation. The input patterns were presented to the networks without any further preprocessing. The frequency range of the input patterns was 200 to 2600 Hz both for the mel-spectra as well as the chop-T representation. Training patterns for the vowel category were taken at aurally detected vowel onsets (see Sec. II A 1) and at 25.6 ms after those onsets. Each training sample was checked visually to exclude erroneous training data like very short vowels or early detections. In all, 1345 training patterns were used: 744 and 601 patterns for the vowel and nonvowel category, respectively. The training patterns for the nonvowel category were chosen in two passes: Initially a small set of nonvowels and silences was chosen. After training, the network was tested on the T-sentence database (see Sec. II A) and patterns that the network erroneously classified as vowels were added to the nonvowel-pattern set. Then, the network was retrained with the vowel set and the extended nonvowel set. The MLP was trained using standard backpropagation with a learning rate of 0.0005 and a momentum term of 0.1. To prevent overtraining we set an error threshold to 0.05. The networks were trained in steps of 100 cycles until the summed error in the vowel versus nonvowel classification no longer declined, usually for 400 to 600 cycles.

3. Vowel-onset detection

Sentences from the training set were processed in steps of 1 ms. The activation of the output unit, representing the likelihood of the presence of a vowel, was first thresholded at 0.6. This threshold value was determined by trial-and-error, and applied to all training and testing conditions. A pseudo-ROC curve is depicted in Fig. 5 for a particular condition, namely, input patterns consisting of 23-channel mel-scaled spectra, with normalization of the rms of the sentence. The ROC curve is determined by evaluation on the PM-sentence database (see Sec. II A).

Like the detection of vowel onsets in the sequence of vowel strength in VSM, we then applied the smoothed derivative filter (see Sec. I A) to trace the local maxima. An example of the output of an MLP as a function of time, based on mel-scaled spectra, is given in Fig. 6 (PM-sentence 13, see Sec. II A).

II. MODEL EVALUATIONS

A. Materials

Hermes (1990) evaluated VSM for a database consisting of 28 Dutch sentences spoken by nine male and nine female speakers, all nonprofessional native speakers of Dutch. They were instructed to read the sentences without taking special care of articulation. This database will be referred to as the T sentences. As described in the Introduction, a trained phonetician traced the vowel onsets by means of the gating technique. The gating technique is based on listening to a short portion of the speech signal by windowing the signal, typically of 20- to 40-ms duration, and shifting this window in

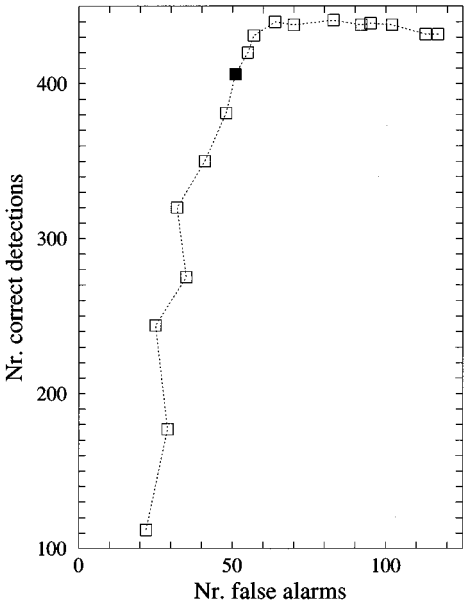

FIG. 5. ROC curve for the threshold parameter of the MLP schemes. The parameter-value range is 0.1 (rightmost data point on the ROC curve) to 0.95 (leftmost point) with a stepsize of 0.05. The filled symbol represents the value of 0.6 as applied in the experiments. These data are obtained for the PM-sentence database (see Sec. II A).

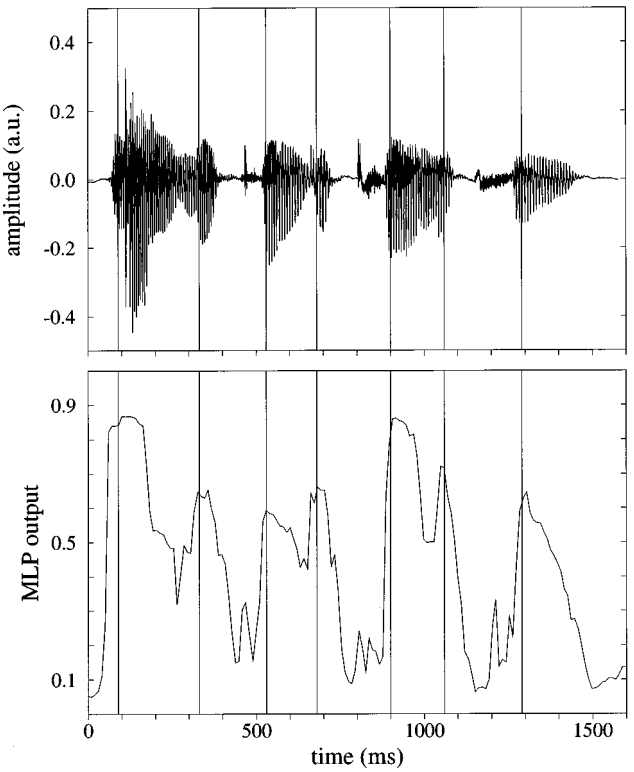

FIG. 6. Waveform of PM-sentence 13 (top) and corresponding MLP output as a function of time (bottom). The MLP is trained on mel-scaled spectra, with normalization of the rms of each sentence. Aurally detected vowel onsets are marked by vertical lines.

TABLE II. Missed-onset and false-alarm rates (as percentages of the number of actual onsets) found for the detection schemes, for the T-sentence database. ''MLP mel-spectra'' refers to training the MLP with mel-scaled spectra, with normalization of the rms of *the whole sentence*. In the case of ''MLP normalized,'' *each individual* mel-scaled spectrum is normalized. ''MLP summed'' refers to the single input unit containing the summed spectral values. ''MLP chop-T'' refers to training with the cochlear-nucleus simulation. See text for a description of the ''channels 1-2-23'' conditions.

		Test level (dB SPL)	Missed onsets (%)			False alarms (%)			Accuracy (%)		
VSM			8			3			91		
CNS		35	16			6			86		
		55	8			9			87		
		75	10			15			84		
CNS–ACF		35	24			3			87		
		55	10			6			88		
		75	11			8			84		
MLP	Mel-spectra		10			10			80		
	Normalized		9			36			67		
	Summed		9			9			77		
			channels			channels			channels		
MLP	Chop-T		1	2	23	1	2	23	1	2	23
	Trained on	35	14	10	12	7	5	3	82	82	85
	test level	55	16	8	10	5	8	9	84	82	82
		75	13	9	8	6	18	14	85	79	77
	Trained on	35	11	14	17	11	15	12	77	74	78
	all levels	55	8	18	14	29	25	24	72	69	73
		75	10	20	8	30	26	24	70	65	73

time through the speech signal ('t Hart and Cohen, 1964). Vowel onsets obtained in this manner are called ''actual onsets,'' 377 of which were detected in the T-sentence database.

Because both the CNS(–ACF) and the MLP schemes have been optimized or trained on the T sentences, an alternative speech database was required for comparison of performances. For this purpose, we made a random selection of 28 out of 560 sentences from the Plomp and Mimpen (1979) set which has its application in diagnostic audiology. We will refer to this selection as the PM sentences. The 28 sentences were rerecorded with 14 male and 14 female nonprofessional speakers resulting in both a male- and a female-speaker version of each sentence. Instructions for reading were similar to those of the T sentences. Actual onsets in these 56 utterances were determined by an experienced phonetician, who traced 466 vowel onsets.

1. Performance evaluation

The number of missed onsets and false alarms are calculated for each of the detection schemes by determining the cross coincidence between actual and algorithmically detected vowel onsets. We adopt the criterion proposed in Hermes (1990) that algorithmically detected onsets should be within ± 60 ms to the actual onsets. Missed-onset and false-alarm rates will be presented as a proportion of the total number of actual onsets. For indication of accuracy, we will also give figures that express the proportion of correct algorithmically detected onsets that are within ± 20 ms to the actual onsets. These results will be presented as ''accuracy'' in the next section. The results for the T sentences have been evaluated for both rates and contexts of occurrences for missed onsets and false alarms. For reasons of clarity, we will only present missed-onset and false-alarm rates. On the basis of these rates, we will select a number of schemes that will be evaluated in more depth for the PM sentences.

B. Results

1. The T sentences

Performance scores for the detection schemes are listed in Table II. The figures given for VSM performance are the same as in Hermes (1990).

For the CNS scheme, a difference between sound-pressure-level conditions can be observed: If all sentences are normalized to 35 dB SPL, the number of missed onsets is about twice the missed-onset rate for 55 or 75 dB SPL. The missed-onset rates for the latter two levels are comparable to VSM performance. The false-alarm rate is substantially lower for the 35 dB SPL condition than for the two other levels. Many more false alarms are obtained with CNS, for the 55 and 75 dB SPL conditions, than with VSM. However, if harmonicity information is used in the detection process, as in CNS–ACF, then the false-alarm rate can be seriously reduced. This effect is most prominent for the 75 dB SPL condition, where analysis of false-alarm occurrences showed that the harmonicity criterion especially rejected onsets for unvoiced plosives. The reduction of false alarms does not affect the missed-onset rate significantly, except for the 35 dB SPL condition. Accuracy figures for the CNS and CNS–ACF schemes are generally lower than the accuracy figure for VSM.

TABLE III. Missed-onset and false-alarm rates (as percentage of the number of actual onsets) found for the detection schemes, for the PM-sentence database. MLP mel-spectra refers to network training with mel-scaled spectra, with normalization of the rms of *the whole sentence*.

		Test level (dB SPL)	Missed onsets (%)	False alarms (%)	Accuracy (%)
VSM			7	9	90
CNS-ACF		35	20	9	92
		55	10	10	91
		75	10	14	90
MLP	mel-spectra		9	14	88

The MLP scheme, with mel-scaled spectra training, yields a missed-onset rate comparable to VSM performance (10%) and a higher false-alarm rate (10%) for the condition where each sentence is normalized. If instead each spectrum is normalized, a comparable missed-onset rate (9%) is obtained while at the same time the false-alarm rate increases dramatically (36%). These results indicate that level differences within a sentence contribute to vowel-onset detection. Accuracy figures for these MLP schemes are also slightly worse than the VSM accuracy figure. An interesting result is obtained for the ''MLP-summed'' condition, which shows comparable missed-onset and false-alarm rates as the mel-spectra case (sentence rms normalized). This finding indicates that instantaneous level or, more specifically, change of instantaneous level contains much information for vowel-onset detection. Analysis of phonemic context of missed onsets, however, reveals that the MLP trained on the full-resolution mel-spectra has a relatively higher missed-onset rate of unaccented /ə/ vowels, and a lower missed-onset rate of accented vowels.

Finally, we will discuss the performance results obtained with MLP networks trained with the simulated chop-T responses, in either the 1, 2, or 23-channel representation. For all representations we applied two training conditions: (a) the network was trained and tested on a single sound-pressure level, and (b) the network was trained on all levels but tested on a single level. In the case of training condition (a) we see that the missed-onset rates are comparable for the 2- and 23-channel representation, and that the false-alarm rate is slightly higher for the 2-channel representation. These results indicate that the 2-channel representation, despite its reduced spectral resolution, contains almost as much information for vowel-onset detection as the 23-channel representation. Comparing the results of this training condition to the CNS model, we find performances which are generally comparable, especially at 55 and 75 dB SPL. The accuracy of the CNS model is overall higher by approximately 5%. If instead only a measure of the instantaneous intensity is used, like in the single-channel condition, the evaluations generally show higher missed-onset and lower false-alarm rates.

For the 2- and 23-channel representations, application of training condition (b) yields an overall trend similar to training condition (a) discussed above. However, missed-onset and false-alarm rates are generally worse than for condition (a): they are higher by a factor of 2 in the case of the 2-channel representation, and moderately higher at low levels in the case of the 23-channel representation. The false-alarm rates are higher by a factor of 2 to 3 for both representations at nearly all levels. These results indicate that the spectral representation in (simulated) chop-T responses is fairly stable over a wide dynamic range as missed-onset rates for the 23-channel representation are comparable for conditions (a) and (b). The increase of missed-onset rate for the 2-channel representation in condition (b) is possibly due to the absence of absolute-level cues within the training set. The same factor probably underlies the increase in false-alarm rate for both representations under condition (b). The results for the single-input condition are in general agreement with this observation; in contrast to the reduction in missed onsets relative to the condition of training and testing on a single level, the false-alarm rate increases dramatically. This finding indicates that the absolute level of (summed) activation is not a very reliable cue.

On the basis of these results, we made a selection of schemes that were tested on the PM-sentence database: the VSM, CNS–ACF, and MLP (mel-scaled spectra training and sentence-level normalization) schemes will be compared. Their performances will be analyzed in more detail in terms of phonemic context of occurrence of missed onsets and false alarms.

2. The PM sentences

Vowel-onset-detection performances for the PM-sentence database are listed in Table III. Missed-onset rates are found to be almost identical to the rates reported above for the T-sentence database. Not only do rates show a high similarity, but phonemic categories of missed onsets correspond as well (see below for a listing of categories for the PM-sentence database). An exception to the latter finding applies to the CNS–ACF scheme: A higher occurrence of missed onsets for /i/ sounds is measured for the PM-sentence database. On the other hand, comparing false-alarm rates for both databases shows that rates for the PM database are generally higher. We suspect that this finding results from differences between both databases in the speakers' care of articulation. Hermes (1990) reported a similar increase of false-alarm rate when testing VSM, optimized for fluent speech, on a database containing carefully articulated, isolated nonsense words. In addition, a contribution to the effect may have come from differences in recording quality, which is

TABLE IV. Occurrences in raw numbers of phonemic contexts of missed onsets, for the PM-sentence database. "CNS–ACF (35)" indicates results obtained for 35 dB SPL input level.

	VSM	CNS–ACF (35)	CNS–ACF (55)	CNS–ACF (75)	MLP
/ə/	17	37	18	19	15
/i/	6	12	9	10	11
/ɪ/	1	13	5	3	5
/e/	4				1
/ɛ/		3	2	2	
/o/	1	2	1	1	1
/ɔ/	1	2	1	2	
/a/	1	4	1	2	3
/ɑ/	1	12	2	1	2
/y/	1	3	2	2	3
/u/	1	1	2	3	3
/ɛɪ/		1	1	1	
/ɸ/		1	1	1	
Σ	34	91	45	47	44

higher for the PM database. As opposed to the performance of VSM and CNS–ACF, the MLP scheme shows a rather stable performance.

An analysis of occurrences of missed onsets is given in Table IV. For all detection schemes main categories of missed onsets are /ə/, /i/, /ɪ/. The VSM scheme does show a moderate number of missed onsets for /e/, but not for the categories /ɑ/, /y/, and /u/. For CNS–ACF and MLP, these latter categories are seen to give rise to a number of missed onsets. Missed onsets for /ə/ sounds generally occurred in unaccented syllables. It is surprising to find that the high vowels /i/, /ɪ/, and to a lesser extent /ɛ/ and /y/, cause difficulties in detection for all schemes. For all schemes, this may be explained by the observation of a delayed detection coinciding with the high second formant reaching its full resonance. Such a detection was typically delayed by more than 60 ms and was thus regarded as a false alarm. We will refer to such phonemic contexts as long vowels.

Table V lists the phonemic categories of occurrences of false alarms. The main general categories are long vowel, /r/, unvoiced-fricative, and nasals contexts. To a slightly lesser extent, false alarms are given for the categories voiced fricatives and /ə/-like. In the case of the latter category, the phonetician did not mark these contexts as vowel onsets because they were poorly articulated. Despite the introduction of a harmonicity criterion in CNS–ACF, we also find a large number of false alarms for unvoiced plosives, even though the false-alarm rate for unvoiced plosives is already largely reduced by this criterion. Also in the case of the MLP scheme, unvoiced plosives and unvoiced fricatives are found to evoke a number of false alarms. With an exception of the unvoiced contexts, the overall distribution of false-alarm categories is in good agreement with the data presented in Hermes (1990; Table I).

III. DISCUSSION

A. Detection schemes

The occurrences of both missed onsets and false alarms in the cases of all three schemes are found to be distributed over more or less the same phonemic categories. This is an interesting finding given the fact that the schemes differ considerably in their preprocessing of the input signals. The main category of missed onsets is "ə-like," occurring in unaccented syllables, which is in agreement with data presented in Hermes (1990). Furthermore, all three schemes have difficulties in detecting onsets of high vowels. In the case of false alarms, a similar distribution is found as presented in Table I in Hermes (1990), although the present evaluations show higher false-alarm rates for /l/ contexts and nasals. As is mentioned in Hermes (1990), some categories of false alarms can function as vowels in other phonetic contexts,

TABLE V. Occurrences in raw numbers of phonemic contexts of false alarms, for the PM-sentence database. "CNS–ACF (35)" indicates results obtained for 35 dB SPL input level.

	VSM	CNS–ACF (35)	CNS–ACF (55)	CNS–ACF (75)	MLP
Long vowel	18	21	16	16	15
/ə/-like	3	3	3	3	4
/r/	4	7	8	8	8
/k/,/p/,/t/	4	1	5	16	15
/l/	2	3	6	5	4
/n/,/m/	5	3	4	5	4
/j/,/w/		1	2	3	1
/b/,/d/				2	1
/χ/,/s/	4		1	3	10
/z/,/v/	2	1	2	4	1
others				2	
Σ	42	40	47	67	63

e.g., in the context of syllabic consonants or in other languages. The fact that all three schemes show similar trends may therefore support the conclusion that the presence of these false alarms has a phonological background.

Within the scope of improving the automated vowel-onset detection, it is found that an MLP network trained with mel-scaled spectra, with normalization of the rms of the whole sentence, is competitive with the VSM scheme. In terms of improvement of the automatic vowel-onset detection, the CNS−ACF scheme does not seem to be an obvious candidate; missed-onset and false-alarm rates are generally higher than for VSM, and the performance of the scheme is found to depend on level. However, this may have a perceptual counterpart. For speech, a comfortable loudness level appears to be well specified. The determination of actual onsets with the gating technique was normally done at a comfortable loudness level, which presumably was substantially higher than 35, and moderately lower than 75 dB sound-pressure level. If the determination had been done at other than comfortable levels, the correspondence between the detection behavior of CNS(−ACF) and the human listener may have been substantially better. In this respect, the performance of the CNS(−ACF) scheme presumably is more comparable to human vowel-onset detection.

It should be noted that the determination of vowel onsets by the human listener was done under conditions with high signal-to-noise-ratio (SNR) levels, where ''noise'' refers to background noises. Likewise, the models were trained and tested with speech signals having high and very high SNR levels for the T and PM-sentence databases, respectively. This means that model performances may deteriorate if trained and/or tested under worse SNR levels. The experiments reported in this paper have not been repeated for such levels, however. Ainsworth and Meyer (1994) investigated speech *recognition* performances using hidden Markov models (HMMs) trained and tested on a number of simulated-response representations of various stages of auditory processing, namely the auditory nerve and cochlear nucleus. Training and testing was done for SNR levels up to 0 dB. Recognition performances were compared to human recognition scores. Of all simulated representations, it was found that the results obtained by using chop-T representations were most like human performance. This finding may implicate that the simulated representation of chop-T responses also provides a reasonably robust representation for vowel-onset detection under low SNR conditions. It should be stressed, though, that this is merely speculative at present and requires experimental evidence.

B. Signal characteristics

The comparative tests give more insight into the relative importance for vowel-onset detection of the signal-characteristics intensity, spectral envelope, and harmonicity.

1. Intensity

The importance of intensity may be derived from comparing vowel-onset-detection models and training conditions. First, for MLP networks trained on mel-scaled spectra, comparable performance is obtained for the conditions of (a) a single input, i.e., the sum of all 23 spectral points, and (b) the full-resolution 23-channel input. In both training conditions, the rms of each sentence processed is normalized. This finding indicates that increments of intensity are important for vowel-onset detection. Second, for the condition that each individual spectrum is normalized before being processed by an MLP network, detection performance is moderate compared to the condition that the rms of each sentence is normalized. Third, if the MLP networks are trained with chop-T representations at all levels and tested on a single sound-pressure level, then the false-alarm rate is substantially higher than with training on a single level. Finally, the difference between the MLP scheme trained on the two-channel chop-T representation and the CNS scheme is that in the latter the derivative of activation plays a key role: Vowel-onset candidates are selected at local maxima of the derivative and its time course is explicitly used as a criterion. Comparing the results for the MLP and CNS, it is found that performances are comparable for the condition that the MLP is trained and tested on a single level. With training and testing on all levels, however, the MLP performance is seen to be substantially worse. Assuming that training of the MLP constructed an optimal classification boundary based on the rough 2-channel spectral envelope, this finding indicates that the derivative of activation can be conceived of as a more robust source of information than spectral envelope *per se*. In sum, the comparison of different models and training conditions indicates that increments of intensity relative to sentence level are of prime importance for modeling human vowel-onset detection. This observation is in agreement with the fact that also in VSM, intensity is an important source of information.

2. Spectral envelope

An MLP network trained with the 2-channel chop-T representation is seen to exhibit reasonable performance, provided training and testing is done on a single sound-pressure level only. A similar, rather crude spectral weighing is the basis of the CNS(-ACF) schemes, which also show reasonable performance. These findings give support to the hypothesis that vowel-onset detection predominantly relies on the rough spectral envelope. On the other hand, false-alarm rates for the CNS−ACF scheme prove to be substantially lower than for CNS, especially at higher sound-pressure levels. This finding indicates that rough spectral envelope is sufficient for detecting vowel-onset candidates, but that more information is required for correct rejection of false alarms. It should be noted, though, that hardly anything is known about integration of activity over such large bandwidths as in the CNS(−ACF) scheme.

3. Harmonicity

The false-alarm rate obtained with the CNS−ACF scheme is substantially lower than the rate obtained with the CNS scheme, while the missed-onset rate is practically identical. This finding may give support to the hypothesis that harmonicity information plays an important role in modeling vowel-onset detection. On the other hand, mel-scaled spectra

do not show a harmonic structure, yet the corresponding MLP performance is found to be satisfactory. This may mean that (a) harmonicity information is present in the mel-scaled spectra in some other way, or that (b) in human vowel-onset detection, harmonicity information is only used to reject some onsets other than vowel onsets. Possibility (a) can be verified by adding harmonicity information to the MLP input, and comparing performance with and without such an extra input unit. Possibility (b) would be in line with the decrease of false-alarm rate observed for CNS–ACF relative to CNS.

C. Onsets in (simulated) chop-T responses

To our knowledge, there are no physiological data reported in the literature pointing at the specific coding of transitions such as vowel onsets. This means that the perceptual information used for the specific detection of vowel onsets cannot directly be derived from measured or simulated response patterns.

Initially, we measured the presence of a formant structure (''vowel strength'') in the simulated chop-T representation. In this array of simulated neuron responses, the formant structure is enhanced by lateral inhibition resulting in a stronger spectral contrast. Blackburn *et al.* (1990) observed that spectral contrast is preserved over a wide dynamic range. However, these measurements were obtained by taking the long-term average of discharge activity during steady state, which may not adequately describe discharge activity at the onset. If spectral contrast is preserved over a broad dynamic range, then we can also expect vowel strength to be stable over a broad range. The measurements did not provide support for this expectation, as especially at high signal levels (75 dB SPL), contrast in vowel strength diminished between vowel and nonvowel contexts. As a result, the false-alarm rate increased considerably, which indicated that the relative contribution of onsets other than vowel onsets started to increase.

D. Syllabification

In this study, a vowel onset was considered to be correctly detected if the detection algorithm signaled a vowel onset within 60 ms of a vowel onset indicated by a phonetician. This rather strong demand can be loosened by considering vowel-onset-detection algorithms as a means for syllabification. For syllabification it is necessary to have one onset per syllable, but this onset does not necessarily have to coincide with the vowel onset. Hunt (1993) presents a study of syllabification by detecting onsets and offsets of vowels by means of recurrent neural networks. The following results were reported: 6% missed syllables and 5% falsely detected syllables, with an accuracy figure for vowel-onset and offset detection of 87%. For the purpose of comparison, Table VI presents missed-onset and false-alarm rates if the three schemes of the present study are applied to detect syllables, instead of vowel onsets. These rates were found by taking one algorithmically detected vowel onset per syllable, while allowing that the onset does not coincide with an actual vowel onset. If there are more than one detections per syllable, these are regarded as falsely detected syllables.

TABLE VI. Rates of *missed syllables* and *falsely detected syllables* (as percentage of the number of actual vowel onsets) found for the detection schemes, for the PM-sentence database. MLP mel-spectra refers to network training with mel-scaled spectra, with normalization of the rms of *the whole sentence.*

		Test level (dB SPL)	Missed syllables (%)	Falsely detected syllables (%)
VSM			5	8
CNS–ACF		35	16	5
		55	7	8
		75	7	12
MLP	mel-spectra		7	11

In comparison to vowel-onset detection (Table III), the CNS–ACF scheme benefits most from this redefinition of the task: In general, the missed-onset and false-alarm rates decrease with approximately 4%. Analysis of the phonetic contexts showed that this improvement especially results from correct syllabification in long vowel contexts, where this context is often seen to cause false alarms in vowel-onset detection. The proportion of falsely detected syllables still is slightly higher than the data presented by Hunt (1993). This may, however, also have resulted from the use of another database (i.e., TIMIT), analogously to the differences found in the false-alarm rate for the T and PM databases in the present study.

E. Perceptual mechanism and phonological structure

It is well known that speech can induce a rhythmic beat which runs synchronously with the syllables which make up the speech signal. Various early studies have shown that the moment of occurrence of this rhythmic beat is close to the vowel onset of the syllable (Rapp, 1971; Allen, 1972; Cole and Scott, 1973). Allen (1972, p. 72) mentions that ''the rhythmic beats were closely associated with the onsets of the nuclear vowel of the stressed syllables, but precede those vowel onsets by an amount positively correlated with the length of the initial consonant(s) of the syllable.'' Cole and Scott (1973) reported the importance of ''vowel transitions'' for the perception of the temporal order of syllables. Studies like these into the rhythmic structure of speech has led to the concept of the perceptual moment of occurrence of the syllable, or its P-center (Morton *et al.*, 1976). Marcus (1981) showed that the length of the onset and the rhyme of the syllable affects its P-center, so that the P-center cannot exactly be identified with the vowel onset. Pompino-Marshall (1989, 1990) developed a detailed model for algorithmic P-center determination, in which he estimates the P-center from the acoustic waveform on the basis of a weighted average of onsets and offsets in the course of the specific loudness of the speech sounds. Since the strongest onsets in a syllable occur in the neighborhood of the vowel onset, where the oral cavity opens and the formants start to rise, this model predicts that the P-center will be close to the vowel onset. In the Pompino-Marshall model, secondary onsets and

offsets in the course of the syllable will move the estimated P-center forward or backward. For a number of syllables, this model is able to predict the shift of the P-center quite accurately. The shifts are never larger than at most a few dozens of ms, however. Therefore, the vowel onset appears to be a good first approximation for the location of the P-center. Also for the rhythmic production of syllables, it is concluded from various studies that, among other phonetic events, the vowel onset best corresponds with the moment speakers use in timing their syllables (Eriksson, 1991).

Phonologically, the syllable has been divided into onset and rhyme. Although this binary division has been supplemented by a division of the rhyme into nucleus (or peak) and coda (Pike, 1947; Selkirk, 1982), it is generally assumed that the boundary between onset and rhyme is more important (Treiman, 1986). In metrical phonology, the syllable is divided into a weak onset and a strong rhyme. The latter in its turn is then divided into a strong nucleus and a weak coda. This implicates that the strongest transition is from the weak onset to the strong rhyme. The transition from nucleus to coda is of a lower level; the transition between two syllables, though of a higher level, takes place from the weak coda of the first syllable to the weak onset of the second. The latter transition is subject to many coarticulatory phenomena, even if the two syllables belong to different words. Therefore, the syllable onset, the transition from one syllable to the next, is much less well defined than the transition from onset to rhyme. In this respect it is interesting to note that the three algorithms tested, though very different in nature, show errors in the same phonetic contexts, which, as argued above, might have a phonological background.

From both a phonetic and a phonological point of view, it is therefore concluded that the best candidate for timing the syllable is the transition from onset to rhyme. In production and perception, the actually realized and perceived temporal moment of occurrence is closely linked with the vowel onset, though higher-order processes can apparently shift the actual moment of occurrence for a few tens of ms. Large shifts occur only when the syllable onset contains (sonorant) segments with strong onsets. For isolated syllables, it may eventually not be too difficult to estimate these shifts from the acoustic speech signal preceding and following the vowel onset. But also in this case, one has to start from the vowel onset. In running speech, the situation is much worse. The syllable onset is often badly defined, in any case much worse than the vowel onset. It is as yet impossible to estimate how much onsets and rhymes will contribute to the rhythmic beat perceived. It is a well known fact that perceived rhythmicity, even in poetry and music, does often not correspond with regular intervals between any well known acoustic event of the sound signal. Nevertheless, we are very sensitive for small changes in the temporal structure of speech signals. In general, it can be concluded that higher-order, ''top-down'' processes push the speech events into a rhythmic frame in which the perception of rhythmicity occurs before it is clearly present in the temporal series of acoustic events. (For a review, see Eriksson, 1991.) In this situation, it is likely that vowel onsets do not precisely correspond with the perceptual moment of occurrence of the syllable. Taking all arguments together, however, they come out as the first phonetic events to be considered in studying perceived rhythmicity in speech.

IV. CONCLUSIONS

In this paper, three methods for automatic detection of vowel onsets in speech have been presented and evaluated. One of these methods, namely vowel-strength measurement, was presented in an earlier publication. For evaluation, two databases of read Dutch speech uttered by nonprofessional male and female speakers have been used. Both databases can be characterized as having a good recording quality with high signal-to-noise ratios. For all methods and corresponding parameter settings, missed-onset rates are found to be better than 25%, and false-alarm rates are not seen to exceed 30%. For the best performing schemes, missed-onset and false-alarm rates are found to be in the order of 10%. In spite of the substantial differences between the three methods, an analysis of the phonetic contexts of missed onsets and false alarms has shown that these contexts generally match.

The method of vowel-strength measurement has been found to be overall best performing. This method is not only based on spectral-envelope information, but also depends to a great extent on both intensity as well as harmonicity information.

The second method presented is based on simulation of chop-T responses found in the anteroventral cochlear nucleus. Reasonably good results were obtained by using a detection scheme that postprocesses the simulated responses, where rough spectral-envelope, intensity, and harmonicity information are used. However, this method was seen to be rather input-level dependent which may have, as has been argued in Sec. III, a perceptual counterpart.

The third detection method is based on training multilayer perceptrons having a single hidden layer. Best results for these MLPs are obtained if the input to the network consists of mel-scaled spectra. Intensity information, i.e., the distribution of intensities over the whole sentence, has been found to be an important source of information in this method. Training the MLPs with the simulated chop-T responses only resulted in fair performance if training and testing was done at the same input signal sound-pressure level.

In summary, the performance results for the different methods and parameter settings support the hypothesis that the main information required for automatic vowel-onset detection are (a) rough spectral envelope and (b) intensity. Harmonicity information can be conceived of as an additional source of information to reduce the false-alarm rate.

ACKNOWLEDGMENTS

The authors would like to thank Armin Kohlraush and Steven van de Par for critically reading the manuscript. Also, the authors thank the two anonymous reviewers for their critical reading and suggestions for improvement. Finally, the first author would like to thank Professor Thomas N. Huckin for providing many recommendations for the process of technical writing. Part of this work was done during a stay

of the first author at Keele University, which was made possible thanks to financial support of the Dutch Research Foundation (NWO).

[1]In Hermes (1990, p. 868) it is stated that ''it is necessary to suppress the contribution of the unvoiced parts. This is achieved by weighing the result of the measurement of the combined strength of the spectral peaks with the maximum value of the subharmonic sum spectrum.'' This phrasing gives the impression that this feature was only applied to avoid some false detections in noisy speech segments. It appears now, however, that this feature was essential for the good performance of the algorithm.

[2]The equation reads:

$$\mathrm{TH}(c_f) = 4.0758c_f^{-1} + 17.4741 - 45.2252c_f + 45.7596c_f^2 - 19.5892c_f^3 + 4.1071c_f^4 - 0.4133c_f^5 + 0.0159c_f^6,$$

where c_f denotes center frequency in kHz, and $\mathrm{TH}(c_f)$ denotes absolute hearing threshold in dB SPL.

Ainsworth, W. A., and Meyer, G. F. (**1994**). ''Recognition of plosive syllables in noise: Comparison of an auditory model with human performance,'' J. Acoust. Soc. Am. **96**, 687–694.

Allen, G. A. (**1972**). ''The location of rhythmic stress beats in English: An experimental study 1,'' Language Speech **15**, 72–100.

Blackburn, C. C., and Sachs, M. B. (**1990**). ''The representation of the steady-state vowel sound /ε/ in the discharge patterns of cat anteroventral cochlear nucleus neurons,'' J. Neurophys. **63**, 1191–1212.

Brown, A. G. (**1991**). *Nerve Cells and Nervous Systems* (Springer-Verlag, London).

Carney, L. H., and Geisler, C. D. (**1986**). ''A temporal analysis of auditory nerve fiber responses to spoken stop consonant-vowel syllables,'' J. Acoust. Soc. Am. **79**, 1896–1914.

Cole, R. A., and Scott, B. (**1973**). ''Perception of temporal order in speech. The role of vowel transitions,'' Can. J. Psychol. **27**, 441–449.

Compernolle, D. S. J. van (**1991**). ''Development of a computational auditory model,'' IPO report 784, Institute for Perception Research, Eindhoven.

Darling, A. M. (**1991**). ''Properties and implementation of the gammatone filter: a tutorial,'' in *Speech, Hearing and Language. Work in Progress* (Dept. Phonetics and Linguistics, University of London), Vol. 5, pp. 43–61.

Darwin, C. J. (**1984**). ''Perceiving vowels in the presence of another sound: constraints on formant perception,'' J. Acoust. Soc. Am. **76**, 1636–1647.

De Boer, E. (**1969**). ''Reverse correlation II: initiation of nerve impulses in the inner ear,'' Proceedings of the Koninklÿke Nederlandse Academie van de Wetenschappen, Vol. 72, pp. 129–151.

Delgutte, B., and Kiang, N. Y. S. (**1984a**). ''Speech coding in the auditory nerve. I: Vowel-like sounds,'' J. Acoust. Soc. Am. **75**, 866–878.

Delgutte, B., and Kiang, N. Y. S. (**1984b**). ''Speech coding in the auditory nerve. II: Processing schemes for vowel-like sounds,'' J. Acoust. Soc. Am. **75**, 879–886.

Delgutte, B., and Kiang, N. Y. S. (**1984c**). ''Speech coding in the auditory nerve. III: Voiceless fricative consonants,'' J. Acoust. Soc. Am. **75**, 887–896.

Eggermont, J. J. (**1985**). ''Peripheral auditory adaptation and fatigue: a model oriented review,'' Hear. Res. **18**, 57–71.

Eriksson A. (**1991**). ''Aspects of Swedish speech rhythm,'' Ph.D. thesis, University of Göteborg, Sweden.

Evans, E. F., and Palmer, A. R. (**1980**). ''Relationship between the dynamic range of cochlear nerve fibres and their spontaneous activity,'' Exp. Brain Res. **40**, 115–118.

Fay, R. R. (**1988**). *Hearing in Vertebrates: A Psychophysics Data Book* (Hill-Fay Assoc., Winnetka), pp. 327–328.

Furui, S. (**1986**). ''On the role of spectral transition for speech perception,'' J. Acoust. Soc. Am. **80**, 1016–1025.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Hart, H. 't, and Cohen, S. (**1964**). ''Gating techniques as an aid in speech analysis,'' Language Speech **7**, 22–39.

Hart, H. 't, and Cohen, S. (**1973**). ''Intonation by rule: a perceptual quest,'' J. Phonetics **1**, 309–327.

Hart, H. 't, and Collier, R. (**1975**). ''Integrating different levels of intonation analysis,'' J. Phon. **3**, 235–255.

Hermes, D. J. (**1990**). ''Vowel-onset detection,'' J. Acoust. Soc. Am. **87**, 866–873.

House, D. (**1990**). *Tonal Perception in Speech* (Lund U.P., Lund).

Hunt, A. (**1993**). ''Recurrent neural networks for syllabification,'' Speech Commun. **13**, 323–332.

Kaufholz, P. A. P. (**1992**). ''Improvement of the Vowel-Onset-Detection algorithm in the IPO intonation meter,'' IPO report 870, Institute for Perception Research, Eindhoven.

Kewley-Port, D., Pisoni, D. B., and Studdert-Kennedy, M. (**1983**). ''Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants,'' J. Acoust. Soc. Am. **73**, 1779–1793.

Kortekaas, R. W. L., and Meyer, G. F. (**1994**). ''Vowel-onset detection using models of the auditory periphery and the nucleus cochlearis: physiological background,'' IPO report 963, Institute for Perception Research, Eindhoven.

Liberman, M. C. (**1978**). ''Auditory nerve response from cats raised in a low noise chamber,'' J. Acoust. Soc. Am. **63**, 442–455.

Marcus, S. (**1981**). ''Acoustic determinants of perceptual centre (P-center) location,'' Percept. Psychophys. **30**, 247–256.

Markowitz, J. (**1993**). ''Listening with intelligence,'' AI Expert **8**, 38–45.

McCulloch, N., and Ainsworth, W. A. (**1988**). ''Speaker independent vowel recognition using multi-layer perceptrons,'' in Proceedings of the 7th FASE Symposium, Vol. 8, pp. 851–858.

Meddis, R. (**1986**). ''Simulation of mechanical to neural transduction in the auditory receptor,'' J. Acoust. Soc. Am. **79**, 702–711.

Meddis, R. (**1988**). ''Simulation of auditory-neural transduction: further studies,'' J. Acoust. Soc. Am. **83**, 1056–1063.

Meddis, R., and Hewitt, M. J. (**1991**). ''Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification,'' J. Acoust. Soc. Am. **89**, 2866–2882.

Meyer, G. F. (**1993a**). ''Models of neurones in the ventral cochlear nucleus: signal processing and speech recognition,'' unpublished Ph.D. thesis, Dept. Communication and Neuroscience, University of Keele.

Meyer, G. F. (**1993b**). ''CNET—point neurone simulator,'' Tech. Rep. TR93-01, Dept. Computer Science, University of Keele.

Morton, J., Marcus, S. M., and Frankish, C. R. (**1976**). ''Perceptual centers (P-centers),'' Psychol. Rev. **83**, 405–408.

Nossair, Z. B., and Zahorian, S. A. (**1991**). ''Dynamical spectral features as acoustic correlates for the initial stop consonant,'' J. Acoust. Soc. Am. **89**, 2978–2991.

Pike, K. L. (**1947**). *Phonemics* (University of Michigan, Ann Arbor, MI).

Plomp, R., and Mimpen, A. M. (**1979**). ''Improving the reliability of testing the speech reception threshold for sentences,'' Audiology **18**, 43–52.

Plomp, R. (**1984**). ''Perception of speech as a modulated signal,'' in *Proceedings of the 10th International Congress of Phonetic Sciences*, edited by M. P. R. van de Broecke and A. Cohen (Foris, Dordrecht), pp. 29–40.

Pompino-Marshall, B. (**1989**). ''On the psychoacoustic nature of the P-center phenomenon,'' J. Phon. **17**, 175–192.

Pompino-Marshall, B. (**1990**). *Die Silbenprosodie. Ein elementarer Aspekt der Wahrnehmung vor Sprachrhytmus und Sprechtempo*, Linguistische Arbeiten 247 (Max Niemeyer Verlag, Tuebingen).

Rapp, K. (**1971**). ''A study of syllable timing,'' in Quarterly Progress and Status Report, Speech Transmission Laboratory, STL-QPRS 1/1971, Stockholm, Sweden, pp. 14–19.

Rhode, W. S., and Smith, D. H. (**1986**). ''Encoding timing and intensity in the VCN of cat,'' J. Neurophys. **56**, 287–307.

Rhode, W. S., and Greenberg, S. (**1992**). ''Physiology of the cochlear nuclei,'' in *The Mammalian Auditory Pathway: Neurophysiology*, edited by A. N. Popper and R. R. Fay (Springer-Verlag, New York).

Rietmole, P. A. te (**1991**). ''Een algoritme ter bepaling van klinkerinzetten en een algoritme ter bepaling van P-centra,'' IPO report 786, Institute for Perception Research, Eindhoven.

Selkirk, E. (**1982**). ''The syllable,'' in *The Structure of Phonological Representation, Part 2*, edited by H. van der Hulst and N. Smith (Foris, Dordrecht), pp. 337–383.

Smith, R. L., and Zwislocki, J. J. (**1975**). ''Short-term adaptation and incremental responses of single auditory-nerve fibers,'' Biol. Cybernet. **17**, 169–182.

Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). ''Dynamic specification of coarticulated vowels,'' J. Acoust. Soc. Am. **74**, 695–705.

Summerfield, Q., and Culling, J. F. (**1992**). ''Auditory segregation of competing voices: absence of effects of FM or AM coherence,'' Philos. Trans. R. Soc. London Ser. B **336**, 357–366.
Tekieli, M. E., and Cullinan, W. L. (**1979**). ''The perception of temporally segmented vowels and consonant-vowels in syllables,'' J. Speech Hear. Disord. **22**, 103–121.
Treiman, R. (**1986**). ''The division between onsets and rhymes in English syllables,'' J. Memory Lang. **25**, 476–491.