Vowel-onset detection

Citation for published version (APA):

Hermes, D. J. (1990). Vowel-onset detection. *Journal of the Acoustical Society of America*, 87, 866-873.
<https://doi.org/10.1121/1.398896>

DOI:

[10.1121/1.398896](https://doi.org/10.1121/1.398896)

Document status and date:

Published: 01/01/1990

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Vowel-onset detection

Dik J. Hermes

Institute for Perception Research/IPO, P. O. Box 513, NL 5600 MB Eindhoven, The Netherlands

(Received 30 January 1989; accepted for publication 30 August 1989)

An algorithm is presented that correctly detects the large majority of vowel onsets in fluent speech. The algorithm is based on the simple assumption that vowel onsets are characterized by the appearance of rapidly increasing resonance peaks in the amplitude spectrum.

Application to carefully articulated, isolated words results in a high number of false alarms, predominantly before consonants that can function as vowels in a different context such as another language or as a syllabic consonant. After applying some modifications in the setting of some parameters, this number of false alarms for isolated words can be reduced significantly, without the risk of a large number of missed detections. The temporal accuracy of the algorithm is better than 20 ms. This accuracy is determined with respect to the perceptual moment of occurrence of a vowel onset as determined by a phonetician.

PACS numbers: 43.72.Ar, 43.72.Lc, 43.70.Fq

INTRODUCTION

The need to detect vowel onsets automatically became apparent in the course of a project aimed at designing a display to provide hearing-impaired persons with visual feedback of the intonation contour of their speech. In order to provide a sentence with a correct intonation, a hearing-impaired person must learn not only to produce the correct pitch movements, but also to position these pitch movements correctly in the syllables in which they occur. If this is not properly done, a pitch movement intended to accent some syllable may fail to do so. Furthermore, some pitch movements can lend prominence to a syllable, but this perceptual effect depends upon the position of the pitch movement with respect to the vowel onset of the syllable ('t Hart and Cohen, 1973; 't Hart and Collier, 1975). Consequently, a full description of accent-lending pitch movements in a sentence requires not only that one knows the course of pitch, but also the moments of occurrence of the vowel onsets.

Vowel onsets are defined here perceptually. Their actual moment of occurrence can be determined by a gating technique described by 't Hart and Cohen (1964). In this technique, short segments of speech are gated out and listened to in isolation. Such a segment may first include only the speech sound preceding the vowel, but when the gate is slowly shifted into the direction of the vowel, a moment occurs at which one clearly starts to perceive the onset of the vowel. It appears that vowel onsets can in most cases be determined with an accuracy better than 20 ms by means of this technique. Voiced fricatives, nasals, liquids, and semivowels, which may function as vowels in other languages, are consonants in Dutch. Hence, their onsets are by definition not considered as vowel onsets. For clarity, vowel onsets should not be confused with other onsets such as voicing onsets or the onsets of the syllables.

Besides from intonation research, the importance of vowel onsets follows from results in other fields of speech perception. Many authors have emphasized the importance of "transients" in general (Cooper *et al.*, 1952; Chistovich *et*

al., 1975; Dorman *et al.*, 1977; Tekieli and Cullinan, 1979; Stevens and Blumstein, 1981; Kewley-Port, 1983; Kewley-Port and Luce, 1984; Strange *et al.*, 1983; Studdert-Kennedy, 1985; Furui, 1986). Kewley-Port *et al.* (1983) concluded that all information about the place of articulation in the transition of a stop consonant to a vowel is confined to a speech segment as short as 20–40 ms. Tekieli *et al.* (1979) demonstrated that the first 10–30 ms of the vowels of V and CV words contain enough information to allow the vowel to be perceived correctly. Furui (1986, p. 1016) showed that "a speech wave of approximately 10 ms in duration that includes the maximum spectral transition position bears the most important information for consonant and syllable perception." Significantly, nearly all these investigations specifically concentrated on vowel onsets such as transitions between stop consonants and vowels, and not on other transients such as voicing onsets, onsets of consonants, offsets, or transitions between consonants. They show that vowel onsets contain information of prime importance for the correct identification of prime importance for the correct identification of the syllable.

The automatic detection of vowel onsets has hardly attracted any specific attention, however. One approach might be to adopt some methods from automatic speech recognition, like, e.g., the one used by Glass and Zue (1987) who start with the detection of the instants in the signal that are characterized by a rapidly changing spectral content. This segmentation is followed by an analysis of the segments. The segmental boundaries followed by a vowel would then naturally produce the vowel onsets. In some other speech-recognition approaches, the vowel onsets might come out as a by-product of a broad phonetic transcription. This often starts with detecting the syllabic nuclei, defined as the maxima in curves that approximate the temporal course of the perceived loudness of the sentence (Mermelstein, 1975; Zwicker *et al.*, 1979). The minima of this loudness function between the syllabic nuclei serve as syllabic boundaries. The "vowellike" parts of the syllables are then found by stripping off the consonant parts (Weinstein *et al.*, 1975; Medress *et al.*, 1978). In these approaches vowel onsets result as sec-

ondary by-products of speech analysis to which no specific significance is attributed. Furthermore, no results are known that verify whether the segment boundaries found by one of these methods concur with any perceptually defined speech events.

The approach presented in this study is more direct. It is assumed that vowel onsets are characterized by a physical characteristic that lends itself to an independent detection. In this paper, this characteristic is the occurrence of rapid increments in the peaks of the amplitude spectrum. It is inspired by some findings of neurophysiological studies of the auditory system. In this perspective, vowel onsets are assumed to correspond with speech events that induce a strong short-term adaptation (Smith and Zwislocki, 1975) in various frequency bands of the peripheral auditory nervous system. Some parameter values used in the algorithm are based on this speculation. We will first describe this procedure and present the results of a test in which its performance is evaluated when applied to fluent speech. Additionally, some modifications will be presented that adjust the algorithm to application to isolated words. Arguments will be given to justify these modifications. Finally, the implications for speech analysis will be indicated.

I. AUTOMATIC DETECTION OF VOWEL ONSETS

The automatic detection procedure for the vowel onsets as proposed here takes place in two stages. In the first, the combined strength of the spectral peaks below 2500 Hz is measured every 10 ms. The result of this measurement will be called "vowel strength" and will be defined more precisely later on. The second stage consists of detecting the instants in the speech signal where the vowel strength rises rapidly. These instants are assumed to be the vowel onsets.

A. Measurement of vowel strength

The procedure for detecting vowel onsets is based on successive measurements representing the strengths of the spectral peaks. The spacing between two such measurements is 10 ms. Each measurement is based on the amplitude spectrum of a speech interval that, for voiced speech, has the duration of one pitch period. This interval starts half the duration of one pitch period before and ends half the duration of one pitch period after the moment for which the vowel strength is calculated. The duration of one pitch period is determined by subharmonic summation (Hermes, 1988), which is also applied in the absence of voicing. Hence, the pitch-determination algorithm is applied every 10 ms, yielding a frequency and a corresponding speech interval, which, for voiced speech, correspond to the pitch and the pitch period of the speech signal at that moment. In the absence of voicing, the resulting period is not related to any perceived pitch, and the duration of the signal from which the vowel strength is calculated is in fact arbitrary.

The measurement of vowel strength at four moments spaced 10 ms is illustrated in Fig. 1, which shows the speech segment around the vowel onset in the Dutch word for "milk," /melk/. The original speech signal is shown in Fig. 1(a). Figure 1(b) shows four speech intervals lasting exactly one pitch period. These periods, as measured by subharmonic

FIG. 1. The measurement of vowel strength around the vowel onset in the Dutch word for milk, melk, /melk/. The original speech signal sampled at 5000 kHz is presented in (a). The arrow points to the moment of occurrence of the physical vowel onset. In (b) four speech segments, spaced 10 ms, each with the duration of a pitch period, are shown. They are resampled at 64 equidistant points. The discrete Fourier transform is presented in (c) after preemphasis and attenuation of components with frequencies below 500 and above 2000 Hz (B_n in text). Observe that the abscissa is presented logarithmically. In (d) the result (C_n in text) is shown of the discrete convolution of B_n with a Gaussian function with a width of one-third of an octave. This convolution is carried out on the logarithmic abscissa. Determination of the vowel strength is illustrated in (e) for the third speech segment. The lengths of the arrows are added up, which, after weighting with maximum of subharmonic sum spectrum, yields the vowel strength.

monic summation, lasted 4.5, 4.7, 5.1, and 5.3 ms, respectively. Observe that, for this speech segment, the durations of these pitch periods are shorter than the 10-ms interval between two successive measurements, so that the four speech segments, by means of which the vowel strength is calculated, do not overlap. In speech in which the pitch is lower than 100 Hz, these intervals do overlap. Each speech interval with the duration of one pitch period is then first resampled at 64 equidistant points. These sample values are calculated by interpolation from the original speech sampled at 5 kHz. The discrete amplitude spectrum of one pitch period is then calculated by means of a fast Fourier transform. This gives the amplitudes of the first 32 harmonics of one pitch period, A_n , $n = 1, \dots, 32$.

This spectrum A_n is preemphasized, and, in order to accentuate the region of the first and second formants, the low (< 500 Hz) and high (> 2000 Hz) frequencies are attenuated according to

$$B_n = \begin{cases} (np/500) [1 - \cos(\frac{1}{2}\pi np/250)] A_n, & 0 < np < 500, \\ (np/500) A_n, & 500 \leq np \leq 2000, \\ (np/500) \{1 - \cos[\frac{1}{2}\pi(2500 - np)/250]\} A_n, & 2000 < np < 2500, \\ 0, & np \geq 2500, \end{cases} \quad (1)$$

where p (in Hz) is the estimated pitch. As n is the harmonic rank of the spectral component, np is the frequency of a harmonic. The spectral components with frequencies below 500 Hz are attenuated in order to suppress the contribution of the first formant in nasal consonants. Including components with frequencies higher than 2500 Hz results, in some cases, in too large a contribution of noisy, high-frequency components, e.g., in voiced fricatives. The only vowel in which the truncation after 2500 Hz may lead to too strong an attenuation of the second formant is the vowel /i/. Figure 1(c) shows a result of this procedure on a logarithmic frequency abscissa.

In the following step the discrete spectrum B_n is convolved on a logarithmic frequency abscissa with the following Gaussian function $H(s)$:

$$H(s) = \exp(-s^2/2\sigma^2), \quad (2)$$

where s is in octaves, and σ equals 0.2 of an octave. The bandwidth between the -3 -dB points is then equal to one third of an octave, which approximates a critical bandwidth. This convolution is carried out on a logarithmic frequency abscissa, in order to take account of the fact that the peripheral auditory nervous system operates in an approximately logarithmic frequency domain. On the logarithmic abscissa this convolution takes the form

$$C_n = \sum_{k=1}^{\infty} B_k H\left(\log_2 \frac{n}{k}\right). \quad (3)$$

This convolution product C_n is calculated for 12 points per octave. A typical result is shown in Fig. 1(d). The convolution of the discrete spectrum B_n with the Gaussian functions is equivalent to substituting an amplitude-weighted Gaussian function for each spectral line and adding up these functions. When the spectral lines are closer to each other than one third of an octave, these Gaussian functions overlap, and as a consequence of the addition of these overlapping Gaussian functions, the result of this convolution is much higher for the narrowly spaced spectral lines at higher frequencies than for the more widely spaced lines at lower ones. This can be observed in Fig. 1 by comparing, for different frequencies, the relative amplitudes of the spectra displayed in Fig. 1(c) with those displaced in Fig. 1(d). The increment in the spectra is much lower for lower frequencies than for higher frequencies. At the lowest frequencies, where the components are separated by more than one third of an octave, each component contributes almost exclusively in accordance with its own amplitude. In other words, the components that are resolved because they are separated by more than the critical bandwidth maintain their individual strength, whereas the unresolved higher components fuse into peaks. The combined strength of the spectral peaks is then determined by adding the increments in the convolved spectrum over those regions where it is increasing. This is illustrated in Fig. 1(e) for the third of the four pitch periods shown in the previous parts of Fig. 1. The result is a measurement that is assumed to represent the combined strength of the resonance peaks in the amplitude spectrum.

As mentioned above, for voiced speech, the measurements are based on a signal with a duration of one pitch period, as measured by subharmonic summation. For un-

voiced parts, too, this measurement is based on a speech signal with a duration corresponding to the frequency at which the subharmonic sum spectrum attains its maximum. (The frequency at which the subharmonic sum spectrum attains its maximum is the estimated pitch. It is a weight sum of the harmonics that contribute to the pitch.) For some noisy parts of the speech signal with much energy at lower frequencies, the result can be too large, and it is necessary to suppress the contribution of the unvoiced parts. This is achieved by weighting the result of the measurement of the combined strength of the spectral peaks with the maximum value of the subharmonic sum spectrum. This value is generally much lower for noisy sounds than for voiced sounds. Hence, this weighting attenuates the parts of the speech signal that are more or less noisy, because the subharmonic sum spectrum is relatively low there. This measure, the combined strength of the spectral peaks weighted with the maximum amplitude of the subharmonic sum spectrum, will be indicated by the term *vowel strength*.

B. Detection of rapid increments in vowel strength

The vowel strength as defined above is measured every 10 ms. An example is given in Fig. 2(a) for the utterance "Do you want to create a new document?" The next stage in the vowel-onset-detection algorithm consists of determining when the vowel strength shows a rapid increase. This is carried out by applying a simple linear filter that simulates short-term adaptation (cf. Smith and Zwislowski, 1975). Indeed, the sequence of measurements is discretely convolved with a biphasic function $P(t)$. This function, shown in the inset of Fig. 2, is taken to be

$$P(t) = \frac{1}{\tau_1 \sqrt{2\pi}} \exp\left(-\frac{(t-d_1)^2}{2\tau_1^2}\right) - \frac{1}{\tau_2 \sqrt{2\pi}} \exp\left(-\frac{(t+d_2)^2}{2\tau_2^2}\right), \quad (4)$$

where $\tau_1 = 15$ ms, $d_1 = 20$ ms, $\tau_2 = 25$ ms, and $d_2 = 5$ ms. The parameters were chosen so that the time course of the simulated adaptation approximated short-term adaptation. The width of the negative part of $P(t)$, τ_2 , was chosen somewhat longer than the width of the positive part, τ_1 , because the time course of suppression is longer than that of excitation. This filter will be referred to as "adaptation filter." An example of applying this filter is shown in Fig. 2(b), which shows the amplitudes of such a convolution product. Observe that relative maxima occur at instants where the vowel strength rises rapidly.

These maxima are candidates for the estimated vowel onsets. Some of them have to be rejected for one of two reasons. First, two maxima may be very close together with hardly any dip between them so that only one of them can correspond to a vowel onset. For fluent speech, this is tackled by rejecting a maximum if the convolution product attains a higher value within 80 ms without there being a negative region between the higher value and the maximum. Second, in speech fragments of low intensity, local maxima can occur to which no significance should be attributed. The amplitude value of any maximum must, therefore, be larger

(a) vowel strength

(b) vowel strength convolved with adaptation filter

FIG. 2. In (a) the vowel-strength measurements are presented for the sentence, "Do you want to create a new document?" The inset at the left shows the adaptation filter with which the sequence of vowel-strength measurements is discretely convolved. The result is shown in (b). The detected vowel onsets are indicated with arrows.

than 1% of the absolute maximum measured in this way for the complete utterance. An example of the vowel onsets thus detected is shown in Fig. 2, in which the vowel onsets are indicated by arrows.

The parameters of this procedure were heuristically adjusted by trying various values to several fluent sentences in different languages, spoken by different speakers. This optimization was not carried out automatically because different weights were attributed to different kinds of error. For instance, a missed detection in an accentuated syllable was considered much more serious than a missed detection before an unaccentuated schwa. The parameters were fixed on the values yielding the most satisfactory results. These values were used for an independent test of the applicability of the procedure to fluent speech.

II. TEST 1: APPLICABILITY TO FLUENT SPEECH

A. Experimental procedure

In order to test the performance of the automatic vowel-onset-detection algorithm for fluent speech, vowel onsets as determined by the algorithm were compared with the vowel onsets as determined by an experienced phonetician. The test material consisted of 28 Dutch sentences other than those used for the setting of the parameters. They were read by inexperienced native speakers of Dutch. Fourteen sentences were spoken by male speakers and fourteen by female speakers. Nine male and nine female speakers were involved, each of whom spoke one or at most two sentences. They were not asked to take any special care in articulating. Dialectic speech characteristics could often be clearly distinguished. An experienced phonetician was asked to indicate the moments of the vowel onsets in these sentences to an accuracy of 10 ms. He performed this task using the gating technique ('t Hart and Cohen, 1964) described in the Introduction where the perceptual definition of vowel onset was presented. This phonetician indicated 377 vowel onsets, which will be referred to as *actual* vowel onsets.

B. Results

The results of the test are summarized in Fig. 3. It shows the cross-coincidence histogram between the actual vowel onsets and the vowel onsets detected by the algorithm. This histogram shows (in bins of 10 ms) the distribution of the time differences between the actual and the detected vowel onsets. The histogram comprises 346 (92%) of the 377 actual vowel onsets, and so 31 vowel onsets went undetected. There were 11 false detections.

Inspection of the width of the cross-coincidence histogram in Fig. 3 shows that 91% of the detected vowel onsets were determined with a precision of 20 ms or better. In the remaining 9%, the discrepancy between the actual and the detected vowel onset was 30 ms or more. Various reasons can be given for this. First, if a vowel is preceded by a glottal stop, the moment of occurrence of the onset cannot be deter-

FIG. 3. The cross-coincidence histogram of the detected vowel onsets and the vowel onsets as determined by an expert phonetician. Included are 346 coincidences (92%) from a total of 377 vowel onsets. Thus 31 vowel onsets were not detected. There were 11 false detections.

mined unambiguously with the gating technique used in this study. Second, a vowel onset after an /h/, which is voiced in Dutch, and after a semivowel, is also difficult to determine perceptually. Only in three cases, timing discrepancies of 30 ms or more occurred after a plosive.

An examination of the 31 undetected vowel onsets showed that they all occurred in short unaccented syllables. All but three of the vowels of these syllables were schwas. In about half of these cases, the actual occurrence of a vowel was equivocal. Indeed, in fluent speech, there will always be some poorly articulated syllables in which it is difficult to ascertain whether the vowel is phonetically realized or not. For the current application of the vowel-onset-detection algorithm, i.e., automatic analysis of pitch movements, these undetected vowel onsets did not matter.

Among the 11 false detections, there were 7 that preceded an /s/ or a /t/. In these cases, the convolution product showed minor maxima that just exceeded the threshold of 1% of the maximum of the total utterance. Two other false detections preceded an /r/; one occurred in a long /a/ and one in an /m/.

C. Comparison with LPC technique

It may be concluded that, for fluent speech, the large majority of vowel onsets are correctly detected. Most undetected vowel onsets precede poorly articulated schwas. Two aspects of the algorithm proved to be essential. First, the fact that the convolution in Eq. (3) was carried out on a logarithmic frequency abscissa and, second, the fact that the 10-ms spaced measurements of the vowel strength were carried out on the basis of one pitch period. This was concluded from some poor results of a first attempt to measure the strength of formant peaks by means of the coefficients obtained with linear predictive coding (LPC). The parameters representing the amplitude and the width of the peaks in the spectral envelope of the speech signal were derived from ten coefficients, obtained with a common LPC analysis based on a fixed analysis window of 25.6 ms. This approach did not lead to very satisfactory results. There are two reasons for this. First, the spectral envelope as determined with LPC produces the peaks on a linear frequency abscissa. With only five formant peaks, corresponding to the ten LPC coefficients, there is a risk that two narrow, high peaks at low frequencies will merge into one broad, low peak. This caused problems in the onsets of vowels such as /o/ and /u/ (cf. Kewley-Port and Luce, 1984). Second, the LPC coefficients were calculated from speech segments of a fixed duration of 25.6 ms. As a consequence, these segments comprised several pitch periods. Frequencies of formant peaks can change at a rate of up to 40 oct/s as can be seen in second-formant transitions in the word "we," /wi/, e.g., Potter (1947), p. 205. Therefore, rapidly changing formant peaks cannot be represented faithfully, but are reduced to broad, low peaks. This is a particular disadvantage for the detection of vowel onsets, because, at these very onsets, formant frequencies tend to change rapidly. Both problems could have been tackled by calculating more than ten LPC coefficients on the basis of one pitch period. This also may have led to good results.

III. TEST 2: APPLICABILITY TO ISOLATED WORDS

A. Too many false alarms

Applying the vowel-onset-detection algorithm with the parameter setting used for fluent speech to isolated, carefully articulated words resulted in an unacceptable high number of false detections. This problem was further investigated by running the algorithm on a large inventory of isolated Dutch utterances, both words and nonwords. This inventory had been recorded from one male speaker with the aim of preparing diphones for Dutch synthetic speech. This guaranteed that all different kinds of vowel onsets were present in the inventory. The inventory contained 1829 words with a total of 5041 vowel onsets. Though the respectable percentage of 99.4% of all vowel onsets was correctly detected, the percentage of false alarms was 15.6%. In these figures, a vowel onset is considered as missed if the algorithm does not detect a vowel onset within 40 ms of its actual occurrence; a detection is considered false if there is no actual vowel onset within 40 ms.

Table I presents the phonetic contexts under which the errors occurred. The various contexts are not independent. For example, a vowel onset can be missed after an /r/ and before a schwa. In these cases an error was counted only once and listed under the context mentioned first in the table. As one observes, a substantial number of errors occur in the context of the Dutch /r/. The Dutch /r/ is mostly an alveolar or uvular trill, different from the retroflex fricative in English. Though the /r/ is a consonant in Dutch, it is worth mentioning that it can function as a vowel in some other languages, e.g., in the Czech word for finger "prst." The /l/, and the nasals, too, can be phonologically equivalent to vowels in some languages other than Dutch, and Table I shows that before these phonemes, too, false alarms occur sometimes. Furthermore, in Dutch as well as in many other languages, the /l/ and the nasals can function as syllabic consonants. A third important context in which false alarms occur is before semivowels, which are perceived as vowels in another phonetic context.

TABLE I. Phonetic contexts in which the vowel-onset-detection algorithm showed false alarms and missed detections in 1829 Dutch isolated words with the parameter setting for fluent speech. Figures are presented as percentages of the total number of vowel onsets (5041).

False alarms	
before /r/	3.8%
before /l/ and nasals	0.7%
before semivowels	1.2%
in vowels and diphthongs	5.4%
in or before /χ/	1.3%
others	3.5%
total (797 false alarms)	15.8%
Missed detections	
before /Vr/	0.3%
after /r/	0.1%
in double vowels	0.1%
before schwa	0.1%
others	0.2%
total (37 missed detections)	0.7%

These considerations support the view that the false alarms in the context of the Dutch /r/, the /l/, the nasals and the semivowels are caused by "vowellike" properties that are especially salient in carefully articulated words. In fluent speech these properties are much less salient, and the consonantal properties of these phonemes predominate.

Another large category of false alarms are those in vowels and diphthongs. This category complements that of missed detections in double vowels, i.e., a sequence of two vowels. In speech, there is a gradual transition between vowels, diphthongs, and double vowels. In fluent speech, a double vowel is often pronounced as a diphthong, whereas, in isolated words, a diphthong can get the character of a double vowel. An ideal vowel-onset detector should draw the same phonetic-category boundaries as the human listener (see, e.g., Peeters, 1989), but it is clear that for this task the present algorithm must be much more refined.

B. Modifications for isolated words

In the light of these results, it was concluded that application of the vowel-onset-detection algorithm to isolated words required some modification. The problem consisted of the large number of false detections. Hence, the conditions imposed upon the maxima in the convolution product of vowel strength and adaptation filter were strengthened. This was based on the assumption that vowel onsets in isolated words are spaced farther apart, and that syllables in isolated words are better articulated, than in fluent speech. In the first place, in the original parameter setting, a maximum in the convolution product between vowel strength and adaptation filter could only be a vowel onset under the condition that no higher value was attained within 80 ms if at least there was no negative region between this maximum and this higher value. This was changed into the stricter condition that there should be no higher value within 80 ms of a maximum, anyhow, whether or not there was a negative dip between this higher value and the maximum. This already reduced the percentage of false detections from 15.6% to 9.7%, while the percentage of missed detections only rose from 0.6% to 1.1%. Next, the two parameters that specify the two conditions of a maximum in the convolution product were systematically varied in order to obtain a setting with fewer false alarms without increasing the number of missed detections too much. The first parameter, par1, originally set to 80 ms, represented the minimum interval between two successive vowel onsets. The second parameter, par2, was the proportion of the absolute maximum of the complete utterance which a maximum in the convolution product should attain. In the original setting this was 1%. The result of varying these parameters is shown in Table II, which presents the number of missed detections and false alarms for a number of values of par1 and par2. The least number of errors was obtained for the values 120 ms for par1 and 3% for par2, which gave 3.8% missed detections and 1.4% false alarms. These values were used to test the applicability of the algorithm for isolated words. The phonetic contexts in which the remaining errors occurred is shown in Table III. (This new parameter setting was unfit for fluent speech.

TABLE II. Number of missed detections and false alarms in 1829 isolated, well-articulated Dutch words containing 5041 vowel onsets for various values of the parameters par1 and par2.

par2 par1	5%		3%		1%	
	missed	false	missed	false	missed	false
0.15 s	12.9%	0.8%	12.4%	0.9%	12.2%	1.2%
0.12 s	4.5%	1.3%	3.8%	1.4%	3.4%	2.3%
0.10 s	3.2%	2.6%	2.5%	3.1%	2.1%	4.6%
0.08 s	2.1%	6.1%	1.4%	7.2%	1.1%	9.7%

Though there were no false alarms, 29% of all vowel onsets in the sentences used for the former test were missed.)

C. Test

A last test consisted of checking whether this new parameter setting could also be used for an arbitrary other set of isolated words. The algorithm was applied to 1080 British-English nonsense words spoken in isolation by one male speaker. In all, these words contained 3316 vowel onsets. Together, they made up a large inventory of isolated words used for the compilation of the diphones of a British-English synthetic-speech system, and so all different kinds of vowel onsets were included.

The outcome of the test for isolated British-English words yielded 96.7% correct detections, 3.3% missed vowel onsets, and 1.4% false alarms, which is about the same as the score for the Dutch words.

The phonetic contexts in which these errors occurred are shown in Table IV. Differences in presentation with Table III stem from phonetic differences between the two languages. There was another difference in the results obtained with the Dutch words and the English words. When the parameter setting for fluent speech was used, the score for the English words was 98.2% correct detections, 1.2% missed vowel onsets, and 4.6% false alarms, which is much better than for the Dutch words. A fair part of this discrepancy

TABLE III. Phonetic contexts in which the vowel-onset-detection algorithm showed false alarms and missed detections in 1829 Dutch isolated words with the parameter setting adjusted for isolated words. Figures are presented as percentages of the total number of vowel onsets (5041).

False alarms	
before /r/	1.0%
before /l/ and nasals	0.1%
before semivowels	0.0%
in vowels and diphthongs	0.2%
in or before /χ/	0.0%
others	0.2%
total (72 false alarms)	1.4%
Missed detections	
before /Vr/	1.1%
after /r/	0.8%
in double vowels	0.6%
before schwa	0.6%
others	0.7%
total (190 missed detections)	3.8%

TABLE IV. Phonetic contexts in which the vowel-onset-detection algorithm showed false alarms and missed detections in 1080 British-English isolated words with the parameter setting adjusted for isolated words. Figures are presented as percentages of the total number of vowel onsets (3316).

False alarms	
in vowels and diphthongs	0.3%
before liquids /l/ and /r/	0.7%
before semivowels	0.4%
total (46 false alarms)	1.4%
Missed detections	
before schwa	1.5%
before /i/	0.3%
before vowel in /CwV/	0.2%
before vowel in /CjV/	0.2%
before vowel in /CIV/	0.5%
before vowel in /CrV/	0.2%
others	0.4%
total (111 missed detections)	3.3%

ancy can be explained from phonetic differences between Dutch and British-English. As mentioned before, the Dutch trill /r/ provided a context in which many false alarms occurred. Such a trilled /r/ is absent from our inventory of British-English words. Another phoneme which caused errors and is absent from English is the Dutch velar fricative /χ/. Together, these two phonemes caused 32% of all false alarms in the Dutch words. For the rest, the discrepancy may be caused by differences in articulation between the two speakers and differences in recording conditions.

The main conclusion is that the vowel-onset-detection algorithm as developed for fluent speech yields a large number of false alarms when applied to isolated words, especially in Dutch, but less so in English. Many of these false alarms have a phonetic explanation. A large proportion can be avoided by modifying two parameters of the procedure without the risk of inducing very many missed detections.

IV. GENERAL DISCUSSION

The procedure presented in this paper for the detection of vowel onsets is basically an elaboration of the assumption that vowel onsets are perceived at those moments at which strong peaks appear or at which existing peaks become much stronger in the amplitude spectrum. It may seem surprising that such a simple procedure scores so well. The results of neurophysiological studies in the cat's auditory nerve (Chistovich *et al.*, 1982; Sinex and Geisler, 1983; Delgutte and Kiang, 1984) may provide an explanation. Delgutte and Kiang (1984, p. 897) concluded that "the peaks in discharge rate that occur in response to certain rapid changes could be used by the central processor as pointers to regions of the spatio-temporal pattern of auditory-nerve activity that are rich in information about phonetic distinctions." This is completely in line with the results of the investigations cited in the Introduction, which showed that all the information required for the correct identification of a transition between a stop consonant and a vowel is confined to an interval of 10–40 ms. In this matter, Delgutte and Kiang (1984) empha-

sized the role of short-term adaptation in the peripheral auditory system. In higher centers of the auditory nervous system, the specific response of neurons to sound segments with changing spectral contents is even more salient (Møller, 1973; Erulkar, 1975; Goldstein and Abeles, 1975; Koch and Piper, 1979; Pickles, 1982). Many neurons show on-responses here, which means that they only respond to the onset of a stimulus. In this perspective, it is postulated that vowel onsets trigger simultaneous on-responses in several frequency channels of the central auditory system. For fluent speech, the algorithm, which, though inspired by this background, is only a very crude implementation, appears to function quite well already. For carefully articulated, isolated words the algorithm functions quite well, too, if some modifications are applied to the setting of the parameters.

It was argued that vowel onsets are characterized by the appearance of strong spectral peaks in the amplitude spectrum of a pitch period. This property must be global in the sense that it involves a wide range of frequency bands. This is presumed to be essential, because an onset can be preceded, as well as followed, by a wide variety of different speech sounds. Any mechanism that detects these onsets should therefore operate in a wide variety of different contexts. Short-term adaptation, which operates in all frequency channels of the peripheral auditory nervous system, fulfills this condition.

The correct detection of vowel onsets is a prerequisite in various fields of speech analysis. As it is necessary for the automatic transcription of pitch movements, any automatic prosodic analysis should include the detection of vowel onsets. Furthermore, the vowel onset indicates where the perceptual information of a syllable is densest. Any speech-recognition system with a design that simulates the human speech processor should therefore incorporate an analysis of what happens immediately before and during a vowel onset. These considerations demonstrate the usefulness of a reliable detection of vowel onsets. The algorithm presented in this study is a simple attempt to realize this.

ACKNOWLEDGMENTS

This research was financed by the Instituut voor Doven, St-Michielsgestel, The Netherlands. It could not have been possible without the stimulating influence of Sieb Nooteboom, Hans 't Hart, and Lei Willems, who all emphasized the role of vowel onsets in speech perception. I am grateful to them and to Adrian Houtsma and René Collier for critically reading the manuscript. Hans 't Hart is thanked additionally for carrying out the tedious job of determining so many vowel onsets, a job which, in fact, should be done by a computer.

- Chistovich, L. A., Fyodorova, N. A., Lissenko, D. M., and Zhukova, M. G. (1975). "Auditory segmentation of acoustic flow and its possible role in speech processing," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London), pp. 221–231.
- Chistovich, L. A., Lublinskaya, V. V., Malinnikova, T. G., Ogorodnikova, E. A., Stoljarova, E. I., and Zhukov, S. Ja. (1982). "Temporal processing of peripheral auditory patterns," in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granström (Elsevier, New York), pp. 165–180.

- Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (1952). "Some experiments on the perception of synthetic speech sounds," *J. Acoust. Soc. Am.* **24**, 597-606.
- Delgutte, B., and Kiang, N. Y.-S. (1984). "Speech coding in the auditory nerve. IV. Sounds with consonant-like dynamic characteristics," *J. Acoust. Soc. Am.* **75**, 897-907.
- Dorman, M. F., Studdert-Kennedy, M., and Raphael, J. (1977). "Stop-consonant recognition: release bursts and formant transitions as functionally equivalent, context-dependent cues," *Percept. Psychophys.* **22**, 109-122.
- Erulkar, S. D. (1975). "Physiological studies of the inferior colliculus and medial geniculate complex," in *Handbook of Sensory Physiology*, edited by W. D. Keidel and W. D. Neff (Springer, Berlin), Vol. V/2, pp. 145-198.
- Furui, S. (1986). "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.* **80**, 1016-1025.
- Glass, J. R., and Zue, V. W. (1987). "Acoustic segmentation and classification," in Proceedings of the DARPA semi-annual review meeting, San Diego, CA, 24-26 March 1987.
- Goldstein, Jr., M. H., and Abeles, M. (1975). "Single unit activity in the auditory cortex," in *Handbook of Sensory Physiology*, edited by W. D. Keidel and W. D. Neff (Springer, Berlin), Vol. V/2, pp. 199-218.
- Hart, J. 't, and Cohen, A. (1964). "Gating techniques as an aid in speech analysis," *Lang. Speech* **7**, 22-39.
- Hart, J. 't, and Cohen, A. (1973). "Intonation by rule: a perceptual quest," *J. Phon.* **1**, 309-327.
- Hart, J. 't, and Collier, R. (1975). "Integrating different levels of intonation analysis," *J. Phon.* **3**, 235-255.
- Hermes, D. J. (1988). "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Am.* **83**, 257-264.
- Kewley-Port, D. (1983). "Time-varying features as correlates of place of articulation in stop consonants," *J. Acoust. Soc. Am.* **73**, 322-335.
- 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.
- Kewley-Port, D., and Luce, P. A. (1984). "Time varying features of initial stop consonants in auditory running spectra: a first report," *Percept. Psychophys.* **34**, 353-360.
- Koch, R., and Piper, H. M. (1979). "Time segmentation in central analysis of complex signals," in *Hearing Mechanisms and Speech*, edited by O. Creutzfeld, H. Scheich, and Chr. Schreiner (Springer, Berlin), pp. 128-132.
- Medress, M. F., Diller, T. C., Kloker, D. R., Lutton, L. L., Oredson, H. N., and Skinner, T. E. (1978). "An automatic word spotting system for conversational speech," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP-78*, 712-717.
- Mermelstein, P. (1975). "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. Am.* **58**, 880-883.
- Møller, A. R. (1973). *Basic Mechanisms in Hearing* (Academic, London).
- Pickles, J. O. (1982). *An Introduction to the Physiology of Hearing* (Academic, London).
- Peeters, W. J. M. (1989). "Vowels, diphthongs, and vowel clusters. A quantitative dynamic approach through synthesis-by-rule," in Proceedings of the workshop on new methods in dialectology held in Amsterdam, The Netherlands, December 1987, edited by M. E. H. Schouten and P. Th. van Reenen (Foris, Providence), pp. 67-77.
- Potter, R. K., Kopp, G. A., and Green, H. C. (1947). *Visible Speech* (Van Nostrand, New York).
- Sinex, D. G., and Geisler, C. D. (1983). "Responses of auditory-nerve fibers to consonant-vowel syllables," *J. Acoust. Soc. Am.* **73**, 602-615.
- Smith, R. L., and Zwislocki, J. J. (1975). "Short-term adaptation and incremental responses of single auditory-nerve fibers," *Biol. Cybernet.* **17**, 169-182.
- Stevens, K. N., and Blumstein, S. E. (1981). "The search for invariant acoustic correlates of phonetic features," in *Perspectives on the Study of Speech*, edited by P. D. Eiman and J. L. Miller (Erlbaum, Hillsdale, NJ), pp. 1-38.
- Strange, W., Jenkins, J. J., and Johnson, T. L. (1983). "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Am.* **74**, 695-705.
- Studdert-Kennedy, M. (1985). "Perceiving phonetic events," in *Perception and Change*, edited by W. H. Warren, Jr. and R. E. Shaw (Erlbaum, Hillsdale, NJ), pp. 139-156.
- Tekieli, M. E., and Cullinan, W. L. (1979). "The perception of temporally segmented vowels and consonant-vowel syllables," *J. Speech Hear. Res.* **22**, 103-121.
- Weinstein, C. J., McCandless, S. S., Mondehein, L. F., and Zue, R. W. (1975). "A system for acoustic-phonetic analysis of continuous speech," *IEEE Trans. Acoust. Speech Signal Process. ASSP-23*, 54-67.
- Zwicker, E., Terhardt, E., and Paulus, E. (1979). "Automatic speech recognition using psychoacoustic models," *J. Acoust. Soc. Am.* **65**, 487-498.