Timing of complex sounds, such as syllables

IPO, Center for Research on User-System Interaction
PO Box 513, 5600 MB Eindhoven

10.09.1998

Rapport no. 1190

Timing of complex sounds, such as syllables

Elise van den Hoven

Voor akkoord:
Dr.ir. R.J. Beun

Timing of complex sounds, such as syllables

Elise van den Hoven
IPO, Center for Research on User-System Interaction
Eindhoven University of Technology
August 1998

Supervisor: Dr. D. J. Hermes

Contents: **page:**

Introduction

General introduction on speech

Rhythm can be induced by synchronous production of syllables. Syllables are speech sounds produced by the human voice, which is a very versatile sound-producing instrument. It can, for instance, sing, hum, moan, shout, whisper and speak. The latter facility is most used of all and is especially important in everyday communication.
In the next two paragraphs, speech production and language will be considered, respectively, whereas in the last paragraph of this chapter the rationale for this report will be presented.

Speech production

The production of speech sounds can be divided into three sequential processes. The first one concerns air movements necessary for speaking. During this process, the air moves from the lungs through the vocal cords of the vocal folds (see Fig. 1). These vocal cords consist of two strong ribbonlike muscles that can close or open the vocal tract. Many speech sounds, in particular voiced speech sounds, are produced when the vocal cords make the airflow oscillate. Speech can, therefore, be seen as a modified way of breathing. The production of these sounds is the second process, which is called the phonation process.

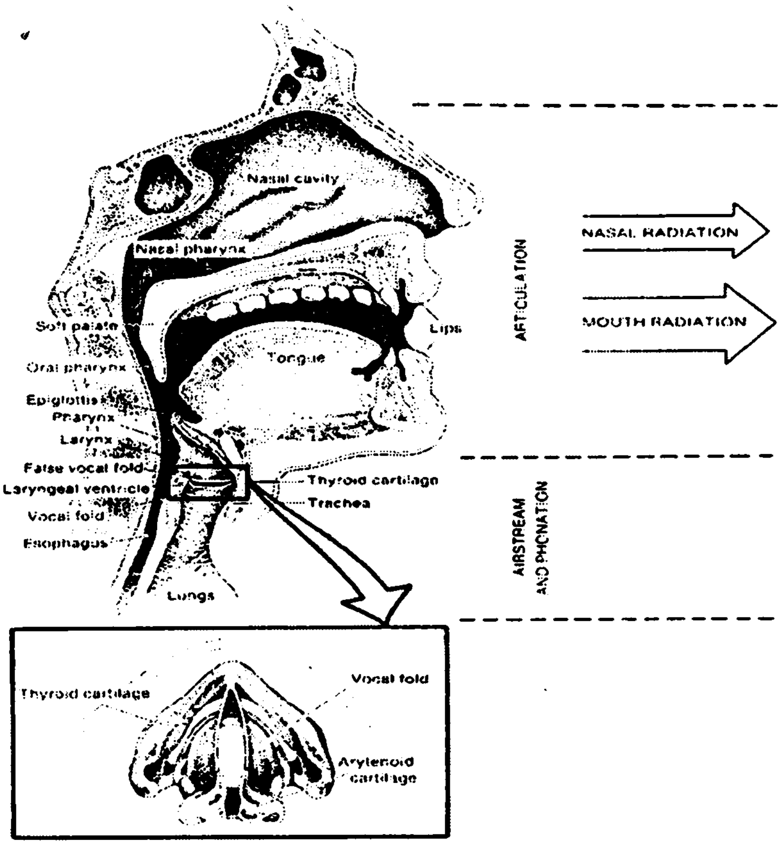

Fig. 1. Illustration of the vocal tract and the vocal cords (Handel, 1989).

Adult male vocal cords are longer and more massive than those of women and of children. Therefore, they have lower fundamental frequencies and thus produce a lower voice (in normal speech, 80-240 Hz for men, 140-500 Hz for women and 170-600 Hz for children, Handel 1989).
The articulation process is the third part in the speech production process. The vibration pattern of the airflow is then affected by the shape of the oral and nasal cavities (see also Fig. 1). This shape is determined most importantly by the tongue, lips, teeth and

palate in the mouth and much less by the relative positions of the lower jaw and the larynx. The reason is that the mouth is more variable in shape and size.
By adjusting the vocal cords and/or the oral and nasal cavities, all kinds of speech sounds can be produced.

Language

‘A six-year-old middle-class American child already recognizes some 13,000 words, while an adult's recognition vocabulary may be well over 100,000’ (Handel, 1989, p. 141). The reason why humans know so many words is because a spoken language is formed only from a small set of basic units. These units, the vowels and the consonants, are the symbols that make up a language when placed in a systematic and meaningful way (Feldman, 1993).

One of the smallest units that makes a difference between speech sounds is called a phoneme and it is written between slashes. In English, for instance, the consonants /b/, /h/ and /m/ are all different phonemes because ‘bug’, ‘hug’ and ‘mug’ have different meanings.
These phonemes have distinctive features. For instance, there are voiced and voiceless phonemes. A voiceless consonant is produced without the vibration of the vocal cords, like the /p/ in ‘pad’. The /b/ is a voiced consonant, because during the production of the /b/, for example in ‘bad’, the vocal cords vibrate. Whether consonants are voiced or not can easily be tested by putting the palm of your hand against your mouth while pronouncing words like ‘bad’ or ‘pad’. Notice the difference in vibrations when saying the /b/ and the /p/.

Another unit of speech is the syllable. No universal definition is known for the term ‘syllable’, but in most cases a vowel (V) is meant preceded and/or followed by a consonant (C) or a group of consonants. For example, the words ‘bug’ and ‘hug’ both consist of only one syllable (a CVC-syllable), whereas the word ‘speaking’ consists of two syllables.
In real utterances, a variable number of syllables is stressed. For example, in the sentence ‘he gave the bear a hug’ the syllables ‘gave’, ‘bear’ and ‘hug’ are stressed.
A term that is also often used in literature on speech research is the word ‘foot’, plural ‘feet’. A foot consists of a stressed syllable followed by a number of unstressed ones.
Schematically a syllable consists of an onset and a rhyme (see Fig. 2 for an example), and the rhyme can again be divided into a nucleus and a coda.

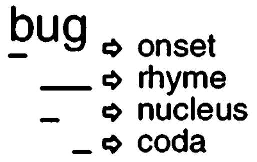

Fig. 2. The schematic division of the syllable ‘bug’ into an onset and a rhyme. In its turn, the rhyme can be divided into a nucleus and a coda.

Why this report?

Speech rhythm is determined by syllables, but it is thus far not exactly known how these syllables are timed. Hence, the aim of this report is to give a survey on research of 'Timing of complex sounds, such as syllables'. Over the past 25 years research has been carried out concerning this topic but no information can be found in standard works concerning speech and hearing such as: 'Persistence and change' edited by Warren and Shaw (1985), 'Listening' written by Handel (1989), 'Auditory scene analysis' written by Bregman (1990) or 'An introduction to the psychology of hearing' written by Moore (1997).

Rhythm in speech

In short, 'rhythm in speech' refers to the perception of an alternation of weak and strong beats (for example 'he GAVE the BEAR a HUG'), but there is no universal definition for rhythm. The definitions known can be divided into two groups, the first is about regularity in time and the second is about regularity of structure in time. For a review, see Eriksson (1991). According to Den Os (1988) there are no theoretical grounds to assume rhythmicity in speech production or in speech perception. (For a recent review on speech rhythm research, see Cummins, 1997.)

According to Pike (1945, p. 34) there are two possible types of 'simple rhythm units' possible for a language, namely stress-timed and syllable-timed languages. (He also distinguishes 'complex rhythm units' in his book, but those will not be discussed in this report.) By a 'rhythm unit' is meant 'a sentence or part of a sentence spoken with a single rush of syllables uninterrupted by a pause' and the rhythm unit is simple because it 'contains only one primary contour'. An example of a simple rhythm unit is: 'The manager is the one who purchased it', where 'man' of the word manager is the only strong stress in the sentence. Among others, Abercrombie (1967), Allen (1972) and Fowler (1979) associated stress-timed and syllable-timed languages with differences in the coordination of the breathing muscles.

According to a review written by Lehiste (1977), some researchers, believe in isochrony whereas others do not, although objective isochrony has never been proved.

Stress-timed rhythm

If a language has a stress-timed rhythm then the interstress intervals are equally long in time, independent on the number of syllables. A term that is used a lot in speech research is isochrony and by this is meant that 'certain linguistic units have the same duration' (Den Os, 1988, p. 7). For example when stressed syllables occur at equal intervals in time they are called isochronous syllables and this is supposed to happen in a stress-timed language. Examples of supposed stress-timed languages are English and Dutch.

Syllable-timed rhythm

This type of rhythm is characterized by equal syllable durations instead of equal interstress interval durations in stress-timed languages, and this results in different kinds of isochrony. A pure isochrony of syllables is thought to be found in syllable-timed languages and isochronous stressed syllables in stress-timed languages. Pike (1945, p. 35) states that in a syllable-timed language 'phrases with extra syllables take proportionately more time, and syllables or vowels are less likely to be shortened and modified'. Examples of supposed syllable-timed languages are French, Spanish and Italian.

A third form of rhythmic timing is mora-timing where a mora is a sub-syllabic rhythmic unit. An example of a mora-timed language is Japanese (for example: Eriksson, 1991).

Den Os (1988) gave a review of all the research done on stress-timing versus syllable-timing and she concluded that these terms 'should not be taken too seriously' (p. 98). Other researchers do not believe in the division of stress versus syllable-timed languages either (Martin, 1972; Dauer, 1983; Strangert, 1985). Dauer (1983) gives three factors that could explain the different rhythms heard in stress-timed or syllable-timed languages: 1) syllable structure, 2) vowel reduction, and 3) phonetic realization of stress and its influence on the linguistic system.

Wenk and Wioland (1982) do believe in stress versus syllable-timed languages but they think specifically that French is not syllable-timed. They state that French has a different kind of stress-timing than English, because the stressed syllables are at the end of the interstress intervals instead at the beginning.

Perceived rhythm

There is also a group of researchers who believe that we can perceive rhythm in speech where it is not present, for instance Lehiste (1977), Fowler (1979) and Benguerel and D'Arcy (1986). Lehiste (1977, p. 262) claims that listeners expect isochrony and use deviations from the pattern to signal syntactic boundaries. Benguerel and D'Arcy (1986, p. 244) think that 'rhythmicity in speech is at the perceptual level, and possibly, at the pre-production level, but not at the acoustic level'. According to Cummins (1997, p. 6), other researchers think rhythm in speech does not exist but is perceived to 'hear temporally patterned events as more rhythmically structured than they actually are (Woodrow, 1951; Fraisse, 1956)'.

Possible functions

The first possible function for rhythm in speech has to do with speech processing. Some researchers think that stress patterns are used for the identification of word boundaries (for example Cutler and Norris, 1988; Cutler and Butterfield, 1992).

Another possibility according to Handel (1989) is that 'speech rhythms aid the listener in segmenting the ongoing acoustic signal into meaningful units but also signify what the speaker is trying to communicate. The speaker imagines what the listener knows or expects, and changes the stress and rhythmic pattern to convey the information'.

A third possible function is that unstressed parts of an auditory pattern are redundant and that only stressed parts are necessary to anticipate on the rest (Sturges and Martin, 1974).

What determines rhythm?

Words that are presented with regular acoustic onsets are in general perceptually not regular. Utterances that contain some kind of rhythm are thus perceptually regular but in general they do not have regular acoustic onsets. Then the question raises: what determines rhythm in speech? Allen (1972) called it syllable beats and Morton et al. (1976) called it P-centers. They both describe the same phenomenon, only Allen tried to determine absolute positions of speech sounds and Morton et al. tried to determine the relative positions of speech sounds.

Timing of syllables

No research has thus far succeeded in finding out where exactly rhythm in speech sounds is located. Therefore a review of the developments is given below.

Syllable beat

One of the first important studies concerning speech rhythm location came from Classe in 1939. Classe tried to find the locations of beats of syllables in English and he did that by letting subjects tap a key to stressed syllables while they read lines of poetry. 'The general result from Classe's experiments was: 'the stress occurs somewhere in the course of the emission of all the consonants considered, with the exception of /b/ and /h/, in the case of which the stress occurs in the course of the following vowel' (p.32) (cited from Eriksson, 1991, p. 14).

Rapp (1971) investigated beat location by instructing subjects to read words in synchrony with pulses. She found that 'pulses are placed earlier in words containing voiceless intervocalic consonants or consonant clusters than in words having a voiced consonant in the same position' (p. 17). Rapp also found a large inter-subject variation in absolute pulse location.

Allen (1972, p. 77) was the first to introduce the term 'syllable beat', because he wanted 'to locate the beat within the syllable'. He tried to do this by tapping and by clicking experiments. During the tapping experiments each subject had to tap to the beat of a syllable heard in an utterance, while the subjects had to match a click to the beat of a syllable in the clicking experiments. The rationale for the clicking experiments was that 'tapping locations are different for different subjects' while 'click-matching [...] does not exhibit inter-subject differences' (p. 189). In spite of the inter-subject differences the subjects placed both clicks and taps before the onset of the nuclear vowels of the stressed syllables by an amount correlated with the length of the preceding consonants. Allen (p. 189) concludes from this that in English 'the rhythmic beats are somehow associated with the consonant-to-vowel transition in the stressed syllables'.
In the same article Allen (p. 190) states that 'a rule for precise prediction of beat location must also include information concerning the articulatory structure of the initial consonant sequence, the degree to which the syllable is stressed, and tempo'. Furthermore, he stated that another 'experiment (Allen, 1968) has shown that the quality of the nuclear vowel and the rhythmic structure of the utterance must be taken into account as well'.
A difficulty concerning the work of Allen (1972) was that a large variability between and within subjects was found. Moreover, others tried to duplicate these experiments in vain: 'workers at IPO in Eindhoven (see van Katwijk, 1974) found different tapping locations for the same subjects on different days' (cited from Marcus, 1976).

P-centers

The psychological moment of occurrence of a word is the definition for the term P-center and was first introduced by Morton et al. (1976). The rationale for this new term was an experiment with digitally stored spoken digits (described in Marcus, 1976). The aim was to make a regular sounding stimulus list containing the English spoken digits ranging from 'one' to 'nine' by using a fixed interval between the digits. But the problem was that the list of stimuli with a regular acoustic onset did not sound perceptually regular. Therefore Morton et al. came up with the term 'P-center' (also spelled as 'P-centre' in British English). Morton et al. found that P-centers do not correspond with word onset, stressed vowel onset or position of peak vowel intensity. This was determined by adjusting the intervals between the nine English-spoken digits and then the relative alignments were studied (for typical exemplars, see Fig. 3). Perceptual centers are assumed to be totally independent of the context.

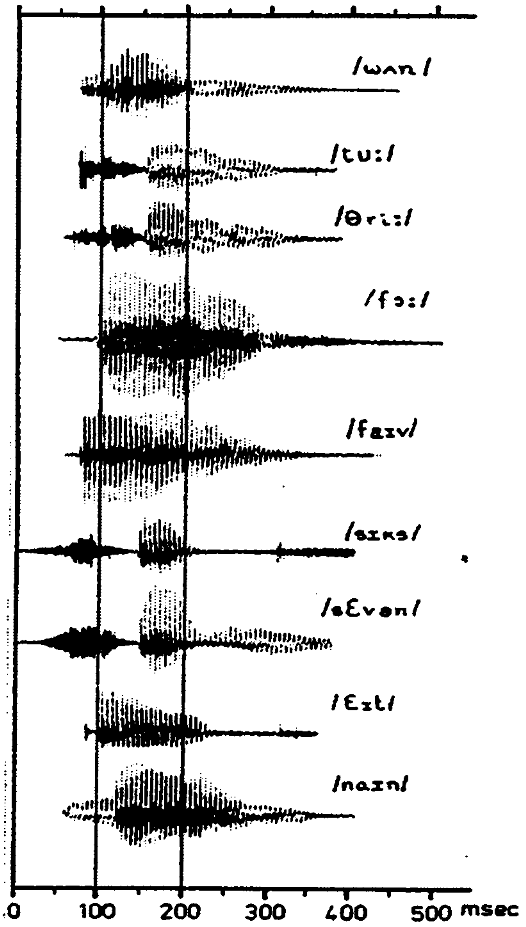

Fig. 3. Amplitude waveforms of the English-spoken digits 'one' through 'nine' (Morton et al., 1976).

Shortly after Morton et al. presented the term 'P-center' in a 'theoretical note', Marcus's thesis (1976) was finished (Marcus was one of the co-authors of the 'theoretical note'). The first experiment he carried out was published in the article of Morton et al., which is described above. This and other experiments were carried out in the following manner. Subjects (in most of these experiments Marcus was the only subject, because results from one experiment showed that there was hardly variability between subjects) had to listen to a pair of syllables, separated by a fixed time interval, that was played repeatedly. Then the subjects had to adjust the position of one of the two syllables to make the cycle sound isochronous. The relative positions of the syllables could then be calculated and thus the P-centers of the syllables were known. This was done for a number of syllable combinations per experiment. (Scott, 1993, called this 'the dynamic rhythm setting task'.)

In one of these experiments, Marcus checked the influence of duration of the initial consonant of a syllable. The results showed that the longer the duration of the initial consonant, the later the P-center after the acoustic onset. The same applied to vowel duration: the longer the vowel, the later the P-center. However, the effect of vowel duration was not as strong as the effect of initial consonant duration. Marcus 'concluded that P-centres were a result of the entire stimulus, rather than one acoustic property of the stimulus' (cited from Scott, 1993, p. 9).

Another experiment concerned modifications in the monosyllabic word 'eight'. Extra gaps, which were added in the word, though hardly audible, did have an effect on the

P-center, whereas amplification of the /t/, which was very salient, had no effect on the P-center.

Marcus (1976) also studied disyllabic words and found that subjects thought of the task as more difficult than with only one syllable. Therefore the results were much more variable. But it was shown that the P-center was not closely related to the stressed syllable. Marcus concluded that 'disyllabic words [...] have two P-centres' (p. 60), one for each syllable.

Also another approach to P-centers was presented by studying dichotic stimuli. From these experiments, which will not be discussed here, he stated that it is logically necessary to define 'simultaneity of dichotic stimuli [...] in terms of simultaneity of P-centres' (p. 133).

The 'two parameter P-centre model' proposed by Marcus will be discussed later. (Some of the above mentioned experiments were published, Marcus, 1981.)

Fowler (1979) related P-centers to articulatory activity instead of acoustical properties. In one of the experiments, one subject had to produce a series of nonsense sentences which were recorded. The sentences, consisting of six monosyllables, were either homogeneously composed (for example: 'mad mad mad mad mad mad') or alternatingly (for example: 'mad sad mad sad mad sad'). All possible combinations were made with the following monosyllables: ad, bad, mad, nad, tad and sad. From the recordings the interstress intervals were measured and it appeared that stress-timed utterances were not acoustically isochronous. The acoustic inter-onset intervals were longer when the consonants of the syllables were of short duration and vice versa.

The conclusion, that stress-timed utterances are not acoustically isochronous, was checked for ten subjects in order to compare production and perception. The original utterances of the experiment described above were used by adding pauses. These were modified in such a way that were acoustically isochronous. Listeners had to indicate whether the natural or the modified version of the utterances sounded more 'rhythmic'. They chose the unmodified version indicating that listeners need deviations from isochrony to 'perceive a sequence as stress-timed' (p. 377).

Two subjects participated in another experiment, for which they had to read sentences which utterances were recorded on tape. The carrier sentence was: 'Jack likes black...' and the words 'acts', 'bats', 'mats', 'gnats', 'racks' or 'sacks' had to make the sentence complete. In short, the result was that 'long intervals are interposed between 'black' and short-duration syllable-initial consonant, while short intervals are interposed between 'black' and a long-duration syllable-initial consonant' (p. 380). Therefore Fowler came up with a simple articulation-based explanation: 'The acoustic phenomena [...] most probably arise in part from differences in the manners of articulation of the consonants and in part from differences in other aspects of their articulatory character' (p. 380).

The last experiment described in Fowler (1979) is about P-center location. Two subjects had to read nine utterances, six homogeneous and three alternating ones. The interstress intervals were measured. The results are described by Fowler as follows: 'In short, the alternating utterances deviate from isochrony in the predicted way, in that intervals starting with prevoiced stops and ending with voiced stops are long relative to intervals that are the reverse of this' (p. 385). Furthermore, she states that 'The experimental outcome supports the articulation-based description of the P-center and fails to support an articulation-free acoustic description' (also p. 385).

A final conclusion drawn by Fowler is that 'due to coarticulation, the articulatory vowel onset would tend to occur during the production of a preceding consonant'. (Coarticulation means that the production of successive speech-sounds overlap.)

Furthermore she hypothesized that 'the acoustic correlates of the P-center may be that (very large) class of acoustic signals that, in the appropriate context, signify (for example) the onset of articulatory activity for a vowel' (p. 386).
Pompino-Marschall (1991) remarks that 'Fowler and coworkers (Fowler, 1979, 1983; Fowler & Tassinary, 1981; Tuller & Fowler, 1980) argue that it might be something like the point of vowel-target near-attainment that articulatorily underlies the P-centre effect' (p. 73).

Eling et al. (1980) repeated two experiments carried out by Marcus (1976), now with Dutch-spoken digits. For the first experiment they found 'good agreement between empirical and predicted values' while for the second experiment the results confirmed 'the independence of the P-centre phenomenon from practice and periodicity' (p. 95). And from that Eling et al. conclude that 'The P-centre concept introduced by Marcus (1976) applies to Dutch digit names in much the same way that it does to English digits' (p. 101).

Hoequist (1983) suggested that P-centers are universal, because he studied a stress-timed language (English), a syllable-timed language (Spanish) and a mora-timed language (Japanese) and found in all three languages the same effect. The English, Spanish and Japanese subjects participating in the experiments had to produce a series of rhythmic utterances. Each utterance was composed of ten alternating monosyllables. The following pairs of monosyllables were used in both orders: *a-ba*, *ma-ba* and *pa-sa*. The interstress intervals were measured and their durations were comparable for all three languages. Hence, the general hypothesis is confirmed that the 'P-center seems to be an aspect of syllable production, and is thus expected to be universal. It is not in itself the mark of a particular timing category' (p. 375).

A different kind of research came from Howell (1984). He studied the amplitude envelope of speech and non-speech signals and concluded that it is an important factor in determining P-center location. From a spoken /ʃa/ a /tʃa/ (short rise) and a /ʃa/ (long rise) were produced in a way that only the envelope was affected. For the non-speech signal A portion of white noise and a portion of a sawtooth waveform were used as the non-speech signal and modified in order to have two different rises, just like for the speech signal. Five subjects participated in each experiment. During such an experiment the subjects had to adjust the intervals between the signals in order to make them sound perceptually regularly (Scott, 1993, called this 'the dynamic rhythm setting task'). The results showed that amplitude envelope alone can give differences in P-center location in speech and non-speech signals. Howell also suggested that Marcus's (1981) manipulations, in which the P-center was varied, could have altered the amplitude envelope as well.

Cooper et al. (1986) carried out an experiment in which they produced a continuum across 'sha' - 'cha' - 'ta'. The ten members of the continuum were presented to the subjects in random order. Three subjects then had to identify the signal as 'sha', 'cha' or 'ta'. The P-center shifted linearly with the duration of the consonant and Cooper et al. thought this to indicate 'that the phonetic identity of syllable-initial consonantal segments does not affect P-center location' (p. 190). This experiment was repeated with another continuum ('sa' -'sta') and the same results were found. Only this time they

interpreted this as disapproving Howell's model (1984), according to Scott (1993, p. 20) 'an incorrect interpretation'.
In a third and fourth experiment Cooper et al. showed that P-centers are not affected by gap duration or the overall duration of the stimulus. 'Instead the P-center appears to have been determined by a combination of at least two different aspects of the signal: the duration of the prevocalic segment and, to a lesser extent, the duration of the vocalic segment' (p. 195).

Fox and Lehiste (1987a), who used the term 'stress beat', although it was considered to be synonymous with the term 'P-center', determined the influence of an addition of an unstressed suffix (such as: *-ing*, or *-able*) or an unstressed prefix (such as: *com-* or *de-*) on the location of the stress-beat of a stressed CVC syllable. They found that an added suffix belated the stress-beat and an added prefix shifted the stress-beat in the opposite direction. The effect of the prefix was much larger than the effect of the suffix.
Other experiments done by Fox and Lehiste (1987b) showed that the duration of the vowel in a CVC syllable determined the stress-beat as well. The longer the vowel duration the later the stress-beat occurred.

Tye-Murray et al. (1987) studied two deaf and two hearing subjects producing rhythmic speech. They had to produce two series of ten monosyllabic utterances, 'tube knock tube knock...' and 'sack splack sack splack...'. Times between different articulatory events were computed and no differences were found between deaf and hearing subjects. Tye-Murray et al. therefore conclude, like Fowler (1979), that P-centers should be considered as articulatory gestures.

The study of effects on P-center location by Cooper et al. which is presented above, is continued in Cooper et al. (1988). This time they studied the effect of the syllable rhyme. In the first experiment they used two continua of stimuli (/a/ and /sa/) in which vowel duration was varied. Three subjects heard syllable pairs (one was a reference syllable, /ba/ and the other one was a member of a continuum) and had to perform a 'dynamic rhythm setting task'. The interpretation of the results was as follows: 'The general increase in the functions indicate that, as expected, the P-center is judged to occur earlier as the duration of the vowel decreases' (p. 234). Because the three subjects showed different results, Scott (1993, p. 21) states in her thesis: 'In the face of such disparate results, it might be suggested that not much can be concluded about vowel duration'.
In the second experiment two different continua were used based on the syllable /at/. In one continuum the vowel duration and thus syllable duration were altered while in the other continuum vowel duration altered in the same but for differences in syllable duration was compensated by adding amounts of silence. The three subjects again had different results. According to Cooper et al. the results suggested 'that the components of the syllable rhyme have equal or near equal effects on P-center location' (p. 239) and 'that the contribution of segments in the rhyme to P-center location varies across listeners in the method-of-adjustment procedure'. According to Scott (1993, p. 22) they should have concluded that the results 'could be explained by other P-centre models, for example the centre of gravity account' (see the model proposed by Howell on page 13 of this report) or 'that the experimental results are too unreliable to base a finding upon'.

Whalen et al. (1989) showed that changing instructions for a P-center-location experiment did not change the results. When three subjects were asked to align pairs of syllables so that 'the syllable onsets, vowel onsets, or syllable offsets sounded isochronous' the results were all the same, indicating that 'syllable timing judgments [...] may, indeed, be the only isochrony judgments that subjects can make reliably' (p. 197).

Pompino-Marschall (1989) presented a psychoacoustic P-center model (which will be described on page 15), after describing three short experiments. In the first experiment the additivity was tested of two linear factors that were varied independently of one another, namely initial consonant duration and vowel duration in CV-syllables. In each trial a syllable, /ma/, was alternated with a click and the click had to be adjusted to sound isochronously with the syllables. From three subjects the results were that 'There was a clear effect of initial consonant duration [...] on P-center location as would be predicted by the two-factor model' (that is the model proposed by Marcus, 1976, see page 13 of this report). The weaker effect of vowel duration was also significant.
The second experiment tested whether the syllable's rhyme affected the P-center location and this was done with the same procedure as described for the first experiment. 'Both effects of vowel and final consonant duration [...] were significant for both subjects' and 'the P-center shifts induced by both vowel and final consonant duration are far more variable than those due to the duration of the initial consonant' (p. 179) were the conclusions.
Whether the P-center effect was only dependent on segment durations and unaffected by the phonetic categorization was tested in experiment three. The methods were the same as for the other experiments, only the syllable used was /ʃi/. Without any exception longer initial consonants resulted in larger P-center shifts, while the vowel effect was only significant for two out of three subjects. 'Taken together, these call for an explanation based more on psychoacoustic parameters than on phonetic phonological segment durations' (p. 182).

Janker and Pompino-Marschall (1991) studied the influence of pitch on the P-center location. One [ka]-syllable was used with five Thai tones, such a syllable was alternated with a click and the two subjects had to adjust the syllable in order to get a subjectively uniform rhythm. Janker and Pompino-Marschall found a significant influence of tone on the P-center location and they conclude that 'in agreement with Schütte (1978) the absolute value of pitch is not important for the position of the P-center but that the temporal changes of the f_0 have a strong effect on its location' (p. 292).

'There is no simple acoustic or articulatory marker of the P-centre position, but [...] nevertheless the P-centre position can be calculated from the acoustic signal' are conclusions by Pompino-Marschall (1990, 1991). These conclusions were based on a series of experiments, some of which will be described below. For one of these experiments two subjects had to utter homogeneous or alternating monosyllables in beat with a metronome. The used monosyllables consisted of /C+ak/, where the consonant could be: /p/, /b/, /f/, /v/, /m/, /ʃp/ or /ʃm/. The results showed that 'the P-centre position is not independent of context but significantly influenced by the nature of the second syllable in the same sequence in almost all of the experiments' (p. 75). Another experiment concerned the influences of differences in syllable rhyme on the P-center location. The subjects had to produce several forms of a German verb, because these forms only differed in rhyme. No influence of syllable rhyme on the P-center location was found.

Articulatory parameters were also studied and the concerning data suggest 'that neither the timing of the consonantal nor the vocalic gesture is a direct correlate of the P-centre and furthermore that the timing of both gestures is not independent of the phonetic composition of the syllable' (p. 83).

Duration variation of phonetic segments of syllables was determined with /ma/-continua. The results showed that 'there was a highly significant effect of the duration of the initial (consonantal) segment for all subjects [...]: Each 40 msec increase in its duration resulted in a significant delay of the P-centre without any exceptions. the weaker effect of the duration of the central (vocalic) segment [...] was not as clear cut' (p. 88).

In the last part of the article Pompino-Marschall presented his psychoacoustic model which will be described in short on page 15 of this report.

Scott and Howell (1992) studied the amplitude/time characteristics of speech signals in order to determine the effects upon P-centers. 'Infinite peak clipping' is a technique which removes the amplitude/time variations of a signal and therefore speech signals ('la', 'wa', 'ra' and 'ya') were infinitely peak clipped. Four subjects had to perform a 'dynamic rhythm setting task' in order to study the differences between original signals and peak clipped signals. 'The P-centers of all speech sounds [...] shifted by the distortion of the amplitude envelope' (p. 154). The experiment was repeated with the four stimuli edited so that they were of the same duration, but the results showed a similar pattern. From the results Scott and Howell concluded that 'These results indicate that durational differences alone are not sufficient to account for differences in P-center location between 'la', 'ya', 'ra' and 'wa'. Envelope and spectral differences must underlie these findings' (p. 155).

Scott (1993) came up with the Frequency dependent Amplitude Increase Model of P-center location, or FAIM, which will be described on page 15. It is a local model based on a series of experiments, all presented in her thesis. These experiments included manipulating the signal's: overall amplitude envelope, stimulus rise time, ramped onset and offset, vowel duration and infinitely peak clipping of the signal.

Scott checked the assumption made by Marcus (1976) that 'the pattern of intervals which lead to the perception of regularity in sequences are the same as the patterns produced when speakers utter isochronous sequences' (p. 71). Eight subjects had to utter repeatedly 'one.. two.. one... two...', which was recorded and afterwards the onset to onset intervals were measured. With a 'dynamic rhythm setting task' the perception of these same utterances was studied. Results of the experiment showed that there was a good correspondence between produced and perceived intervals, which supported Marcus's hypothesis, but there was variance between subjects.

The second experiment concerned differences between speakers because of the variance found in the former experiment. The same utterances were used to investigate whether different P-center locations caused the differences between subjects and this appeared to be the case.

Experiment three was carried out to study the influence of infinite peak clipping on speech signals, but this experiment is described above in Scott and Howell (1992).

The effect of the amplitude envelope on the P-center location was described in experiment four. The /t/ in the word 'eight', produced by a male speaker, was amplified, supposing this would shift the location of the P-center towards the offset of the signal.

But the results of the experiment indicated that the amplification did not affect the P-center either in production or perception.
Stimulus rise time was another variable manipulated to investigate the influences on the location of the P-center. Scott changed the ramping of natural speech in order to manipulate the rise times and by a 'dynamic rhtyhm setting task' the P-center locations were determined. 'Ramping the onset of speech sounds, and thus changing their measured rise times, alters their P-centers. The shift caused by this manipulation is away from the onset of the signal. Longer rise times lead to later P-centres' (p. 162). This experiment was duplicated but then to study the effect of ramped stimulus offsets, but the effect was negligible.
In experiment seven, the effect of vowel duration was tested on P-center location. The results showed that, when amplitude profile and spectral content are held constant, the effect was not significant.

A citation by Scott (1993, p. 222) about context independence of P-centers: 'Lame (1990) suggested that a principle assumption of P-centres - that they are context independent (Morton et al. 1976) is incorrect. P-centres, he argues, like other rhythmic phenomena are influenced by the interval perceptions, and thus by the content of the other signals in the rhythmic sequence. [...] Evidence to support Lame's assertion that P-centre are not context independent comes from Seton (1989); he found a difference on the effect of the intensity of stimuli on their P-centres in mixed or blocked loudness presentations'.

The most recent article known by the author comes from De Jong (1994). De Jong evaluated in two experiments articulatory models of P-center location. He concluded that 'P-center locations do not correspond to any particular kinematic articulatory event, but rather to a complex of events taken from throughout the stimuli' (p. 447). The experiments were performed by four subjects who had to do 'the dynamic rhythm setting task' with the words 'toats' and 'totes', which were edited digitally (more fully described in De Jong, 1991). Also acoustic events (final offset, voicing offset, peak amplitude and voicing onset) and articulatory events (three points on the tongue, dorsal, mid and tip, the upper and lower lip, and two points on the jaw) of the speaker of the stimuli were studied. De Jong concluded that 'The articulatory event that corresponds closest to the subject's responses is the minimum position of the tongue tip (and jaw for 'toast') in the vowel. Acoustically, the most appropriate event is the onset of voicing, although the timing of peak amplitude and offset of voicing also reduce the variance in subject's responses' (p. 451). Furthermore he states that 'the timing of the tongue tip minimum corresponds well to P-center responses, as does the timing of voice onset'.

Summary on P-centers

The psychological moment of occurrence of a word is the definition of the term P-center (Morton et al. 1976) and it is found in syllables. Thus far no concluding combination of factors is known which determine P-center location. There are several good candidates, however, namely:
1) articulatory gestures (for example: Fowler, 1979; Tye-Murray, 1987), coarticulation is proposed as the determining factor;

2) psychoacoustic parameters (for example: Howell, 1984; Pompino-Marschall, 1989, 1990, 1991), like rise time and amplitude envelope;
3) or phonetic / phonological variables (for example: Morton et al., 1976; Marcus, 1976; Cooper, 1986) like initial consonant duration and vowel onset.

Modeling of syllable timing

Modeling of syllable timing is necessary for the prediction or use of speech rhythm for example for talking user interfaces. Therefore some researchers have tried to develop such a model. All models of syllable timing known by the author concern P-center models. They can be divided into three types of models, respectively: models based on phonetics or phonology, models based on articulatory gestures and models based on psychoacoustics.

Models based on phonetics or phonology

The most recent model came from Scott (1993) who also gives an elaborate review on P-center modeling in her thesis. In her review she makes a distinction between global and local models, where a local model maps P-centres onto a single acoustic event and a global model does not. (Her model will be discussed later.)

Marcus (1981) was the first to develop a model for P-center location and he described it by:

$$P\text{-}centre = \alpha x + \beta y + k$$

(see also Fig. 4). Where *P-centre* stands for the P-center location relative to the stimulus onset, x is the initial consonant or consonant cluster, y is the vowel plus the final consonant duration, α and β are parameters of the model and k is an arbitrary constant.

This model was tested by Marcus and he found the values 0.65 and 0.25 to fit the parameters α and β, respectively. Marcus's model fitted his data well. A disadvantage of Marcus's model is that it is only based on the results of one subject. Scott (1993) preferred this model above those of Howell (1988) or Vos and Rasch (1981), because its predictions for the results of her experiments were better.

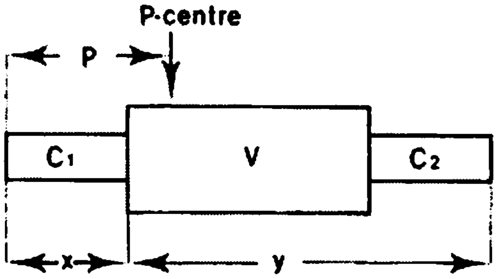

Fig. 4. Schematic representation of the 'two parameter P-centre model' of Marcus (1981).

The second model concerning P-centers came from Howell (1988). He stated that the experiments done by Marcus (1981) altered the speech waveforms and this could have affected the 'behavior' of P-centers, especially because he concluded from his experiments (Howell, 1984) that amplitude envelopes influence P-center location. Therefore his global model contained some parameters concerning the amplitude envelope of the speech signal (see also Fig. 5 for a schematic representation):

$$\textit{center of gravity} = ((-0.5ah)\ (a / 3) + ((l - a)/2)\ (1 - a)h)) / (0.5ah + (l - a)h)$$

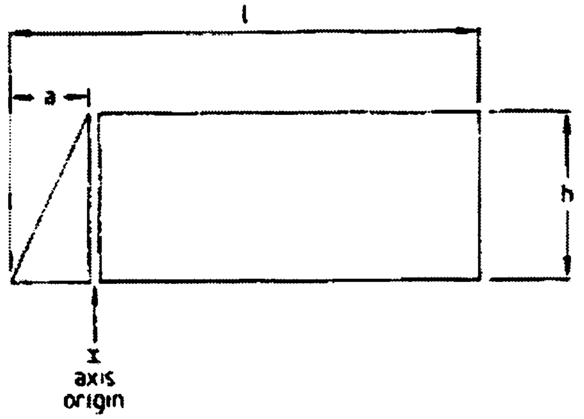

Fig. 5. Amplitude envelope of a syllable as an example for Howell's model (1988).

By *center of gravity* is meant the relation between P-center and the envelope of the syllable. The following parameters: *a*, *h* and *l* represent respectively: 'length of triangle along abscissa', 'height of triangle and rectangle along ordinate' and 'length of the polyhedron along abscissa' (p. 91). According to Scott (1993), a disadvantage of Howell's model is that it is only based on experiments in which only one factor was varied, while a lot of factors are involved in the determination of P-center location. An advantage of Howell's model is that it can be used for speech as well as for non speech signals.

Scott (1993) came up with the Frequency dependent Amplitude Increase Model of P-center location, or FAIM. It is a local model based on a series of experiments. These experiments included manipulating the signal's: peak clipping, overall amplitude envelope, stimulus rise time, ramped onset and offset and vowel duration. The model is based on the 50%-point of the maximum amplitude of the rise time (see Fig. 6 and for more information, see Scott, 1993, chapter twelve), and fitted her data well.

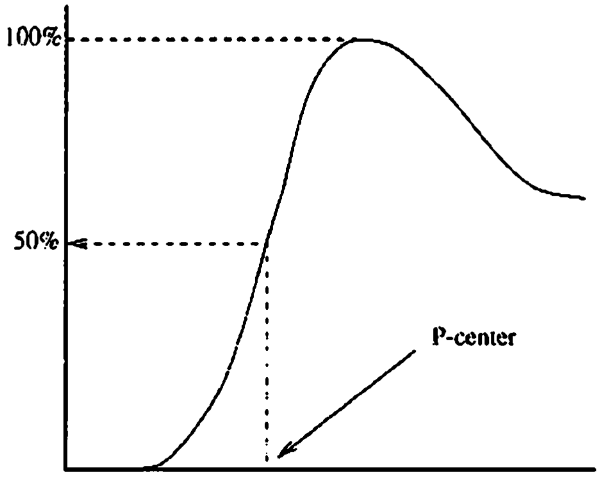

Fig. 6. Determination of the P-center location in the model of Scott (1993) (schematic representation made by Cummins, 1997).

Model based on articulatory gestures

A different kind of P-center model came from Fowler (1979) who related P-centers to articulatory activity (see also page 8). She hypothesized that 'the acoustic correlates of the P-center may be that (very large) class of acoustic signals that, in the appropriate context, signify (for example) the onset of articulatory activity for a vowel' (p. 386). Furthermore, she also stated that 'due to coarticulation, the articulatory vowel onset would tend to occur during the production of a preceding consonant'. Scott (1993) states that this model is as well local as global.

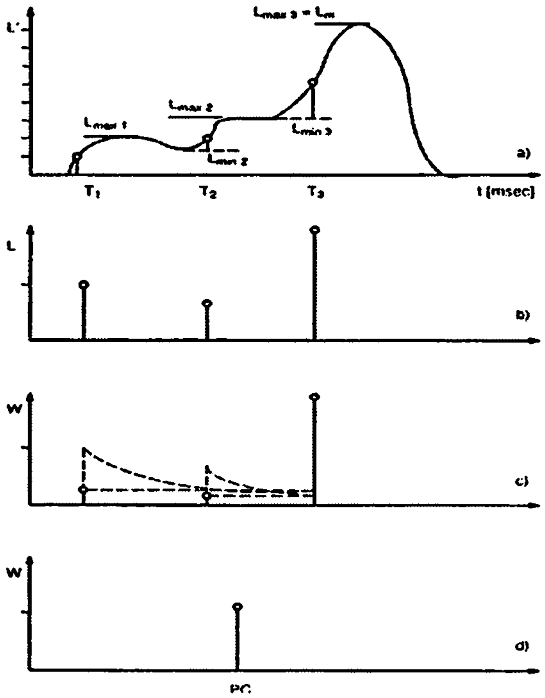

Fig. 7. Four steps during the determination of the P-center, according to the model by Pompino-Marschall (1991).

Model based on psychoacoustics

Pompino-Marschall (1989, 1990, 1991, 1992) proposed a psychoacoustic P-center model (for an exact description see Pompino-Marschall, 1990 and for a review of the four most important steps during the determination of the P-center location, see Fig. 7). In short it 'uses thresholds within loudness functions within each critical band; it also uses temporal weighting processes between these 'partial events' and integration processes at different levels. The model is complex, but can be generally expressed as determining a perceptual syllabic centre of gravity, from an integration of the partial events' (Scott, 1993, p. 24). Pompino-Marschall (1989, p. 185) remarked that 'although the exact weighting factors and the exact type of integration [...] have still to be determined experimentally, it can be shown that the model clearly can cope with all P-center phenomena reported so far'.

Timing of complex sounds other than syllables

How does a dancer move to the beat of the music? And how can a group of musicians play isochronously? Can P-centers or syllable beats be found in music? Thus far there are no concluding answers to these questions, because most of the timing related research is about speech, but the few articles that have been written about timing in complex sounds are described below.

Vos and Rasch (1981) showed that the perceptual onset of alternating tones (modulated sinusoids) of 70 dB can be defined by the point at which the envelope is 15 dB lower than he maximum level of the tones. When the intensity of the alternating tones decreases the relative thresholds increase. They also propose an explanation for their results and that is 'that adaptation of the hearing mechanism to a certain relative stimulus level is responsible for perceptual onset' (Vos and Rasch, 1981, p.334).

According to Gordon (1987) the P-center in speech is closely analogous to the perceptual attack time (PAT) in music and he defines PAT 'as the time a tone's moment of attack or most salient metrical feature is perceived relative to its physical onset' (p. 88). This way, he distinguishes the PAT from the perceptual onset time, just like the P-center does not correspond with the moment at which the signal becomes audible (take for example the monosyllable 'stress', the signal becomes audible when you first hear the /s/, but the P-center is placed later than the /s/). [Gordon also remarks that Vos and Rasch (1981) indicated the PAT with 'perceptual onset'.] From the experiments carried out with 16 different instrument tones, Gordon concludes that the PAT mainly depends on the slope of the rise time function.

Could this account for speech as well?

References

Abercrombie, D. 1967. *Elements of general phonetics.* Edinburgh: Edinburgh University Press.

Allen, G. D. 1968. Experiments on the rhythm of English speech. *Working papers in phonetics,* No. 10 (UCLA), 42.

Allen, G. D. 1972. The location of rhythmic stress beats in English: an experimental study, parts I and II. *Language and speech,* 15 (1), 72-100, 179-195.

Benguerel, A.-P. and D'Arcy, J. 1986. Time-warping and the perception of rhythm in speech. *Journal of phonetics,* 14, 231-246.

Bregman, A. S. 1990. *Auditory scene analysis: The perceptual organization of sound.* Cambridge, Massachusetts: The MIT Press.

Classe, A. 1939. *The rhythm of English prose.* Oxford: Basil Blackwell.

Cooper, A. M., Whalen, D. H. and Fowler, C. A. 1986. P-centers are unaffected by phonetic categorization. *Perception and psychophysics,* 39 (3), 187-196.

Cooper, A. M., Whalen, D. H. and Fowler, C. A. 1988. The syllable's rhyme affects its P-center as a unit. *Journal of phonetics,* 16, 231-241.

Cummins, F. 1997. *Rhythmic coordination in English speech: An experimental study.* PhD thesis, Indiana University.

Cutler, A. and Butterfield, S. 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of memory and language,* 31, 218-236.

Cutler, A. and Norris, D. 1988. The role of strong syllables in segmentation for lexical access. *Journal of experimental psychology: human perception and performance,* 14, 113-121.

Cutting, J. E. and Rosner, B. S. 1974. Categories and boundaries in speech and music. *Perception and psychophysics,* 16, 564-570.

Dauer, R. M. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of phonetics,* 11, 51-62.

De Jong, K. J. 1991. The articulation of consonant-induced vowel duration changes in English. *Phonetica,* 48, 1-17.

De Jong, K. J. 1994. The correlation of P-center adjustments with articulatory and acoustic events. *Perception and psychophysics,* 56 (4), 447-460.

Den Os, E. 1988. *Rhythm and tempo of Dutch and Italian; A contrastive study.* PhD thesis, University of Utrecht.

Echols, C. H., Crowhurst, M. J. and Childers, J. B. 1997. The perception of rhythmic units in speech by infants and adults. *Journal of memory and language*, 36, 202-225.

Eling, P. A., Marschall, J. C. and van Galen, G. P. 1980. Perceptual centres for Dutch digits. *Acta psychologica*, 46, 95-102.

Eriksson, A. 1991. *Aspects of Swedish speech rhythm.* PhD thesis, University of Göteburg.

Feldman, R. S. 1993. *Understanding psychology.* (3rd ed.). New York; McGraw-Hill, Inc.

Fowler, C. A. 1979. "Perceptual centers" in speech production and perception. *Perception and psychophysics*, 25, 375-388.

Fowler, C. A. 1983. Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in sequences of monosyllabic stress feet. *Journal of experimental psychology: General*, 112, 386-412.

Fowler, C. A. and Tassinary, L. 1981. Natural measurement criteria for speech: The anisochrony illusion. In J. Long and A. Baddeley (eds.), *Attention and performance IX.* Hillsdale, New York.

Fox, R. A. and Lehiste, I. 1987a. Effect of unstressed affixes on stress-beat location in speech production and perception. *Perceptual and motor skills*, 65, 35-44.

Fox, R. A. and Lehiste, I. 1987b. The effect of vowel quality variations on stress-beat location. *Journal of phonetics*, 15, 1-13.

Fraisse, P. 1956. *Les structures rhythmique.* Érasme, Paris.

Gordon, J. W. 1987. The perceptual attack time of musical tones. *Journal of the Acoustical Society of America*, 82 (1), 88-105.

Handel, S. 1989. *Listening: An introduction to the perception of auditory events.* Cambridge, Massachusetts: The MIT Press.

Hoequist, C. E. 1983. The P-center and rhythm categories. *Language and speech*, 26 (4), 367-376.

Howell, P. 1984. An acoustic determinant of perceived and produced isochrony. In M. P. R. Van den Broecken & A. Cohen (Ed.), *10th International congress of phonetic sciences* (pp. 429- 433), Dordrecht, Holland: Foris.

Howell, P. 1988. Prediction of P-center location from the distribution of energy in the amplitude envelope: I. *Perception and psychophysics*, 43, 90-93.

Janker, P. M. and Pompino-Marschall, B. 1991. Is the P-center position influenced by 'tone'? *ICPhS*, Aix-en-Provence, 3, 290-293.

Lame, G. D. 1990. *Timing perception and control: A new internal clock model and modality-specific phase resetting.* PhD thesis, University of Texas in Austin.

Lehiste, I. 1977. Isochrony reconsidered. *Journal of phonetics*, 5, 253-263.

Marcus, S. 1976. *Perceptual centers.* PhD thesis, University of Cambridge.

Marcus, S. 1981. Acoustic determinants of perceptual centre (P-centre) location. *Perception and psychophysics*, 30, 247-256.

Martin, J. G. 1972. Rhythmical (hierarchical) versus serial structure in speech and other behavior. *Psychological review*, 79 (6), 487-509.

Moore, B. C. J. 1997. *An introduction to the psychology of hearing.* (4rth ed.). San Diego: Academic Press.

Morton, J., Marcus, S. and Frankish, C. 1976. Perceptual centers (P-centers). *Psychological review*, 83 (5), 405-408.

Pike, K. L. 1945. *The intonation of American English.* University of Michigan Press, Ann Arbor, MI.

Pompino-Marschall, B. 1989. On the psychoacoustic nature of the P-center phenomenon. *Journal of phonetics*, 17 (3), 175-192.

Pompino-Marschall, B. 1990. *Die Silbenprosodie. Ein elementarer Aspekt der Wahrnehmung von Sprachrhythmus und Sprechtempo.* Tübingen: Niemeyer (= Linguistische Arbeiten 247; Habilitationsschrift).

Pompino-Marschall, B. 1991. *The syllable as a prosodic unit and the so-called P-center effect (29).* Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM).

Pompino-Marschall, B. 1992. The P-center and the perception of tempo and rhythm in connected speech. *Fourth workshop on rhythm perception & production*, Bourges, 157-162.

Rapp, K. 1971. A study of syllable timing. *Papers from the institute of linguistics university of Stockholm*, 8, 14-19.

Schütte, H. 1978. Subjektiv gleichmäßiger Rhythmus: Ein Beitrag zur zeitlichen Wahrnehmung von Schallenereignissen. *Acustica*, 41, 197-206.

Scott, S. 1993. *P-centres in speech; an acoustic analysis.* PhD thesis, University College London.

Scott, S. K. and Howell, P. 1992. Infinite peak clipping alters the P-center of speech. *Fourth workshop on rhythm perception & production*, Bourges, 151-156.

Seton, J. C. 1989. *A psychophysical investigation of auditory rhythmic beat perception.* PhD thesis, University of York.

Strangert, E. 1985. *Swedish speech rhythm in a cross-language perspective.* PhD thesis, University of Umeå.

Sturges, P. T. and Martin, J. E. 1974. Rhythmic structure in auditory temporal pattern perception and immediate memory. *Journal of experimental psychology*, 102, 377-383.

Tuller, B. and Fowler, C. A. 1980. Some articulatory correlates of perceptual isochrony. Perception and psychophysics, 27, 277-283.

Tye-Murray, N., Zimmermann, G. N. and Folkins, J. 1987. Movement timing in deaf and hearing speakers: comparison of phonetically heterogeneous syllable strings. *Journal of speech and hearing research*, 30 (3), 411-417.

Van Katwijk, A. F. V. 1974. *Accentuation in Dutch, an experimental linguistic study.* PhD thesis, Utrecht University.

Vos, J. and Rasch, R. 1981. The perceptual onset of musical tones. *Perception and psychophysics*, 29 (4), 323-335.

Warren, Jr. W. H. and Shaw, R. E. 1985. *Persistence and change: Proceedings of the first international conference on event perception.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Wenk & Wioland, '82. Is French really syllable timed? *Journal of phonetics*, 10 (2), 193-216.

Whalen, D. H., Cooper, A. M. and Fowler, C. A. 1989. P-center judgments are generally insensitive to the instructions given. *Phonetica*, 46, 197-203.

Woodrow, H. 1951. Time perception. In Stevens, S. S., (ed), *Handbook of experimental psychology.* Wiley, New York.

Summary

Rhythm in speech, or in other complex sounds, is determined by syllables, but it is not known where exactly. In order to solve this problem, the term ‘P-center’ was introduced. The definition of a P-center is ‘the psychological moment of occurrence of a word’ (Morton et al. 1976, p. 405). After the introduction of this term, a lot of research has been carried out concerning P-centers, but no concluding factors determining P-center location have been found so far. The vast majority of P-center research was carried out with a low number of subjects, ranging from one to three.

Acknowledgments

Dr. D. Hermes, thank you very much for your help and your nice comments. Dik, I enjoyed working with you.

Ir. D. J. Breebaart, Jeroen, thank you very much for your support, ILYSM!