THE PROSODY OF SPEECH: MELODY AND RHYTHM

Sieb Nooteboom
Research Institute for Language and Speech
Utrecht University
Trans 10
3512 JK UTRECHT
Netherlands

1. INTRODUCTION

The word *'prosody'* comes from ancient Greek, where it was used for a "song sung with instrumental music". In later times the word was used for the "science of versification" and the "laws of metre", governing the modulation of the human voice in reading poetry aloud. In modern phonetics the word 'prosody' and its adjectival form 'prosodic' are most often used to refer to those properties of speech that cannot be derived from the segmental sequence of phonemes underlying human utterances. Examples of such properties are the controlled modulation of the voice pitch, the stretching and shrinking of segment and syllable durations, and the intentional fluctuations of overall loudness. On the perceptual level these properties lead amongst other things to perceived patterns of relative syllable prominences, coded in perceived melodical and rhythmical aspects of speech. In modern generative phonology (Selkirk, 1984; Nespor and Vogel, 1986), the word 'prosody' has been given a somewhat different meaning, as it refers to nonsegmental aspects of abstract linguistic structure, such as a particular type of constituent structure and the presence or absence of accents, that are, at least potentially, systematically reflected in the phonetic rendition of utterances. Of course, the phonetic and phonological meanings of the word prosody might be considered two sides of the same coin: although phonologists give primacy to an abstract description of the phenomena concerned, they look for empirical evidence in the realm of speech. Phoneticians rather start from observations on real speech, but the abstract notions they come up with to account for the observed phenomena are phonological by nature. In this chapter we will take our starting position in the phonetic domain.

From a phonetic point of view we observe that human speech cannot be fully characterised as the manifestation of sequences of phonemes, syllables or words. In normal speech we hear for example that pitch moves up and down in some non-random way, providing speech with recognizable melodical properties. We also hear that segments or syllables are shortened or lengthened, apparently in accordance with

some underlying pattern. We hear that some syllables or words are made to sound more prominent than others, that the stream of words is subdivided by the speaker into phrases made up of words that seem to belong together, and that, one level higher up, these phrases can be made to sound as if they relate to each other, or, alternatively, as if they have nothing to do with each other.

Properties of speech that cannot be derived from the underlying sequence of phonemes are often called suprasegmental properties of speech, including whether speakers speak soft or loud, whether they speak in a normal, a hoarse or a breathy voice, whether they articulate carefully or slurringly, or even whether they would speak with an unusual posture of the vocal tract and the larynx so as to disguise their voice. Typically prosodic features of speech are not reflected in normal orthography, nor in conventional segmental phonetic transcriptions. In this chapter the treatment of prosody on the phonetic level will be limited to speaker-controlled aspects of voice pitch, organized in perceived speech melody or intonation, and speaker-controlled aspects of speech timing, organized in the perceived rhythmical structure of speech. Such melodical and rhythmical aspects of speech seem to have a variety of communicative functions, most of these closely tied to the fact that they mediate between the abstract and time-free mental structures underlying speech utterances and the production and perception of speech developing in real time.

Section 2 will discuss the melodical structure of speech, section 3 the rhythmical structure of speech, and in section 4 some communicative functions of speech prosody will be discussed.

2. THE MELODY OF SPEECH

2.1 Introduction

This section deals with speech intonation, in its strict interpretation as "the ensemble of pitch variations in the course of an utterance" ('t Hart, Collier and Cohen, 1990:10), concentrating on those pitch variations that are related to perceived speech melodies, and thereby paying less attention to pitch variations that are related to the segmental structure of speech. As the knowledgeable reader will notice, the subject is approached following the ideas of 't Hart et al. (1990), giving primacy to the perceptual structure of intonation. The reader is referred to their book for a much fuller account and argumentation, based on more than 25 years of intonational research. One should also notice that the insights presented here are claimed to have validity only for so-called intonation languages such as the Germanic languages, Romance languages, and Japanese. Tone languages such as Chinese, in which lexical forms are distinguished by differences in level and/or movement of pitch on a particular vowel phoneme, are not dealt with in this chapter.

Obviously, the approach by 't Hart et al. is not the only one in the world. Other attempts to come to grips with the structure of intonation in terms of the actual course of pitch in speech utterances and its perceptual and linguistic correlates, can be found in Fujisaki and Sudo (1971) for Japanese, Maeda (1976), O'Shaughnessy (1976; 1979), Pierrehumbert (1980) for American English, Brown, Currie and Kenworthy (1980) for British English, Bruce (1977) for Swedish, and Thorsen (1980; 1985) for Danish. What these approaches all have in common is that they strive for some kind of stylized approximation of the apparently capricious pitch fluctuations found in natural speech, hence making reality more tractable by data reduction. In the approach by 't Hart et al. it is demonstrated that one can find a reliable basis for such stylization in the way pitch contours are perceived by native listeners, and that intonation can be described in terms of sequences of standard discrete pitch movements, supposedly corresponding to voluntary actions on the part of the speaker.

2.2 Speech pitch, production and perception

In strict terms, pitch is the perceptual correlate of F_0, the fundamental frequency or repetition frequency of a sound. One should be aware, however, that rather often the notion "pitch" is used to refer to F_0 or the repetition frequency itself. In speech F_0 is determined by the rate of vibration of the vocal cords located in the larynx. The physiological and acoustic mechanisms by which F_0 is controlled are rather intricate

and will not be dealt with here. An excellent account of these mechanisms is given in Borden and Harris (1983). Rate of vibration of the vocal cords, and thereby F_0, is measured in Hertz (Hz; 1 Hz is 1 cycle per second). The range of F_0 for each individual speaker mainly depends on the length and mass of the vocal cords. For males in conversational speech this range is typically between approximately 80 and 200 Hz, for females between approximately 180 and 400 Hz, and for young children this range can be even considerably higher. Within this range each speaker has to a large extent active control over F_0: a speaker can choose to speak on a high or a low pitch, and can produce pitch rises and falls. However, many details of the actual course of pitch in speech are not actively controlled by the speaker, but are rather involuntary side-effects of other speech processes, often related to the production of particular speech sounds. For example, other things being equal, high vowels like /i/ and /u/ have a higher intrinsic pitch than low vowels like /a/ (Peterson and Barney, 1952; Ladd and Silverman, 1984; Steele, 1986). In vowels following voiceless consonants the voice pitch starts higher than in vowels following voiced consonants (Ohde, 1984; Silverman, 1986). These involuntary aspects of speech pitch superimpose small perturbations on the course of pitch, and often, in the visual analysis of measured pitch fluctuations in speech utterances, make it difficult to identify those pitch variations that are responsible for the perceived speech melody.

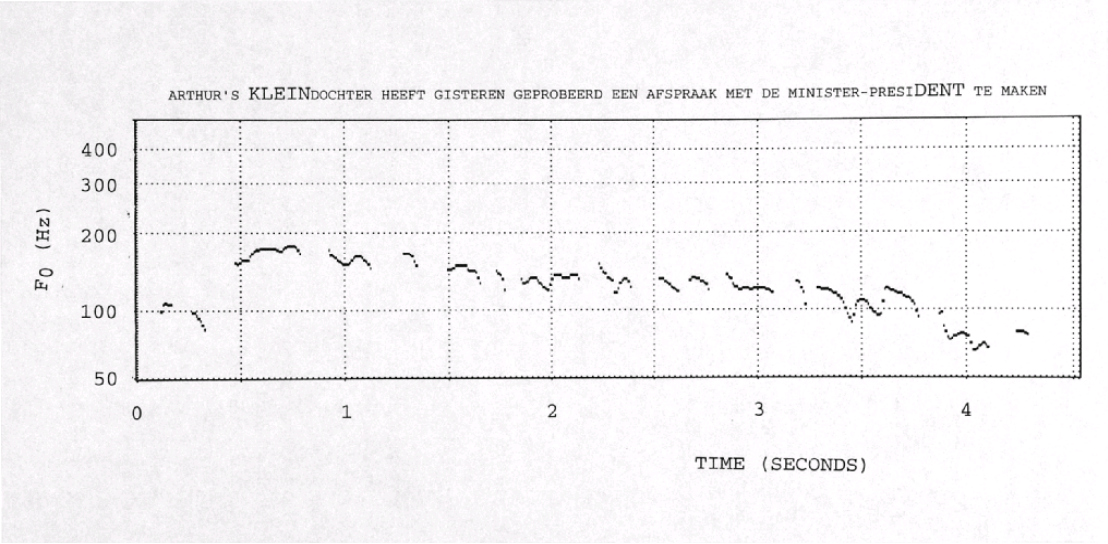

Fig. 1: Measured course of pitch in a Dutch sentence, with only two voluntary pitch movements, an accentuating rise on the syllable “KLEIN” and an accentuating fall on the syllable “DENT”. All other pitch movements are involuntary side-effects of other speech processes. Fig. 1 about here Note also that the continuity of pitch is interrupted during all voiceless consonants.

This is illustrated in Fig. 1, presenting the measured course of pitch in a rather lengthy Dutch utterance. In fact, this sentence was spoken with only two accent-representing pitch movements, one on “klein” in “kleindochter” and one on “dent” in “president”. All the other pitch movements do not seem to contribute to the perceived speech melody. These other pitch movements are thought to be involuntary side-effects of other processes of articulation.

In Fig. 1 it can also be seen that the course of pitch is discontinuous. It is interrupted during the production of voiceless consonants like /k/, /p/, /t/. One should note that, while listening to fluent speech, one does not hear these interruptions of F_0 during stop consonant silent intervals as pauses or perceptual interruptions of the course of pitch. Although such interruptions of F_0 contribute to the perceived character of the consonants concerned, as listeners, we have the illusion that the speech and its intonation or melody are uninterrupted. In fact, interruptions of the sound of speech during for example silent intervals of voiceless consonants are only perceived as interrupting the stream of speech and the speech melody when they are longer than, roughly, 200 ms. This is in accordance with the observation that silent gaps longer than this effectively prohibit perceptual integration of preceding and following speech sounds. Sensory information about the preceding speech sound has faded away during the silent gap and therefore is not available for perceptual integration when the following speech sound arrives (Huggins, 1975; Nooteboom, 1979). It should also be noted that, when the pitch after a silent interval is considerably higher or lower than before, the listener perceives a rise or fall in pitch, as if human perception unconsciously bridges the silent gap by filling in the missing part of the pitch contour. It is only when the virtual pitch change becomes unnaturally steep, that this illusion breaks down and is supplanted by perceptual disintegration of the stream of speech, potentially leading to the perception of a second speaker interrupting the first (Nooteboom, Brokx and De Rooij, 1978).

In quasi-periodic complex sounds like voiced speech, pitch is perceived on the basis of the frequency interval between harmonics present in the signal. We might think of some central processing mechanism, finding the common divisor of a number of candidate-harmonics detected in the signal (Goldstein, 1973). The third to the sixth harmonics are most effective, thereby constituting a dominance region for periodicity pitch (Ritsma, 1967). The lowest harmonic or fundamental (i.e. F_0 itself) does not need to be physically present for pitch to be perceived (Schouten, 1940; note that, if the fundamental were necessary, normal male speech would have no perceivable pitch over the telephone, where frequencies below 300 Hz are generally filtered out).

Human pitch perception is, for signals with clearly defined periodicity, remarkably accurate, the differential threshold (just noticeable difference) being in the order of 0.3 to 0.5 percent (Flanagan and Saslow, 1958). In natural speech accuracy of pitch perception varies considerably as a function of the clarity of periodicity in the signal. This clarity of periodicity covers the whole range from the absence of any periodicity in silent gaps or voiceless fricatives, via ill-defined periodicity in voiced fricatives and in hoarse or overly breathy voices, and during rapid pitch changes, to well defined periodicity in vowels with sufficient loudness, produced with well-vibrating vocal cords and without rapid pitch changes. For this reason, which has to do more with the highly variable nature of human speech than with perceptual acuity, it is not feasible to give a general figure of how accurately pitch in speech can be determined, either by humans or by machines. However, it is safe to assume that during most vowel sounds in normal speech with a normal loudness level, pitch can be determined with an accuracy of a few percent.

For the study of intonation, pitch distances are more relevant than absolute pitch: we can recognize the same melody in different pitch ranges, for example those of a male and a female speaker. For this reason it is often useful to measure pitch in semitones rather than in Hertz, the semitone scale being just one possible log scale derived from the Hertz scale. The distance D in semitones between two frequencies f_1 and f_2 is calculated as:

$$D = 12*\log_2(f_1/f_2) = 12/\log_{10}2*\log_{10}(f_1/f_2)$$

One semitone roughly corresponds to a frequency difference of 6 percent. The reader should be aware, however, that the semitone scale, although adequate for predicting pitch distances, is not adequate for predicting equal perceptual prominences made by pitch movements in different (for example male and female) registers. A psychoacoustic scale derived from the frequency selectivity of the auditory system, associated with distances along the basilar membrane, appears to be more appropriate for this purpose. This psychoacoustic scale can be approximated by the so-called Equivalent Rectangular Bandwidth (ERB) scale, calculated as:

$$E = 16.7 \log_{10}(1+ f/165.4),$$

$$f = 165.4 (10^{0.06E} - 1),$$

in which E is the ERB-rate (number of ERBs corresponding to a particular frequency) and f is frequency in Hz (Hermes and Van Gestel, 1991).

Precisely because in speech perception pitch distance is more relevant than absolute pitch, the differential threshold of pitch distance is more relevant than the differential threshold of pitch itself. It has been estimated that only pitch differences of more than three semitones can be discriminated reliably (‘t Hart, 1981; ‘t Hart, Collier and Cohen, 1990:29). This would suggest that pitch differences smaller than three semitones cannot play a role in speech communication, but Rietveld and Gussenhoven (1985) showed that pitch differences of 1.5 semitones create reliable differences in the perception of prominence.

2.3 Perceptual equality: close-copy stylizations

It has been assumed above that there are many apparently capricious details in the pitch fluctuations in speech utterances that are not actively controlled by the speaker, but are rather involuntary side-effects of other speech production processes. It was also assumed that such involuntary pitch movements do not contribute to the perceived speech melody. A priori this is a bold assumption, comparable to the assumption that involuntary, segmentally conditioned, variations in speech sound durations are irrelevant to the perceived rhythmical structure of speech. As we will see later, the latter assumption does not hold. But the earlier assumption with respect to pitch fluctuations does hold. This can be shown by using the technique of analysis-by-synthesis, replacing the original pitch course of an utterance by an artificial one, using for example an LPC-analysis-resynthesis system (Atal and Hanauer, 1971) or the more recent Pitch Synchronous Overlap and Add method (PSOLA: Hamon, 1988; Charpentier and Moulines, 1989).

A first step in this demonstration is the so-called ‘close-copy stylization’ of pitch in speech utterances, as applied by De Pijper (1983) to British English. Fig. 2 gives the natural, measured F_0 curve of an English utterance together with its close-copy stylization. A *close-copy stylization* is defined as a synthetic approximation of the natural course of pitch, meeting two criteria: it should be perceptually indistinguishable from the original, and it should contain the smallest possible number of straight-line segments with which this perceptual equality can be achieved. Note that the graphical representation of the close-copy stylization continues through the voiceless portions in the utterance. In the actual resynthesis voicing, and therewith pitch, will be suppressed in these voiceless portions.

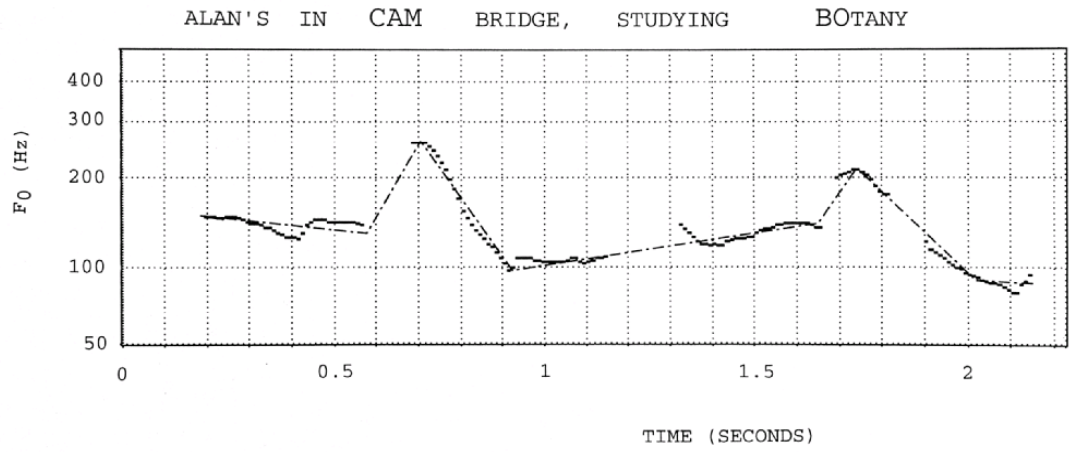

Fig. 2: Measured course of pitch (dotted line) in a British-English utterance together with a so-called "close-copy" stylization (interrupted line), containing the smallest possible number of straight-line segments with which perceptual equality between original and close-copy can be achieved.

De Pijper (1988) has convincingly demonstrated in a formal experiment using 64 native speakers of British-English as his subjects, that the capricious pitch curves of natural utterances can be simplified to sequences of straight-line segments in the time - log F_0 domain, without there being any noticeable difference between original and close-copy pitch curves. This is an important finding, because it justifies the description of intonation in terms of rather simple approximations. One should notice that there is no reason why this finding should be limited to approximations with straight-line segments. If, for example, one were to conduct a similar study using some well-defined curvilinear approximations such as cosine-functions (e.g. Fujisaki and Sudo, 1971), one would obtain a similar outcome: there seems to be no fundamental reason to use straight-line segments in describing intonation. However, as we will see below, there is a practical reason. Straight-line segments easily lend themselves to a description of intonation in terms of neatly segmented discrete units ('t Hart et al., 1990:71).

2.4 Perceptual equivalence: towards standard pitch movements

Close-copy stylization is based on perceptual equality. It is only a first step in the description of intonation. If we have someone imitate the intonation or speech melody of an utterance, either with the same words or with different words, or even with no words at all by humming, we obtain a pitch curve that will definitely not be perceptually equal to the original. It will be easy to hear many differences. But yet we, or a panel of native listeners, can hear whether the imitation is successful in

conveying the same melodic impression. Apparently, intonation is organized in terms of melodic patterns that are recognizable to native speakers of the language. This calls for a unifying notion different from perceptual equality. For this other notion ‘t Hart et al. use the term *perceptual equivalence*. Two different courses of F_0 are perceptually equivalent when they are similar to such an extent that one is judged a successful (melodic) imitation of the other (‘t Hart et al., 1990:47). Perceptual equivalence implies that the same speech melody can be recognized in two realizations despite easily noticeable differences, in the same way that the same word can be recognized from different realizations.

The powerful notion of perceptual equivalence now allows us, for any intonation language, to set up, by various sorts of generalizations, an inventory of standard pitch movements. Combinations of these generate pitch contours that are perceptually equivalent to naturally-occurring pitch curves. Fig. 3 gives a measured pitch curve for a British English utterance, together with its close-copy stylization and a standardized stylization. Notice that, in order to arrive at this standardized stylization, a grid is set up with three equidistant lines in the time - log F_0 domain. These three equidistant

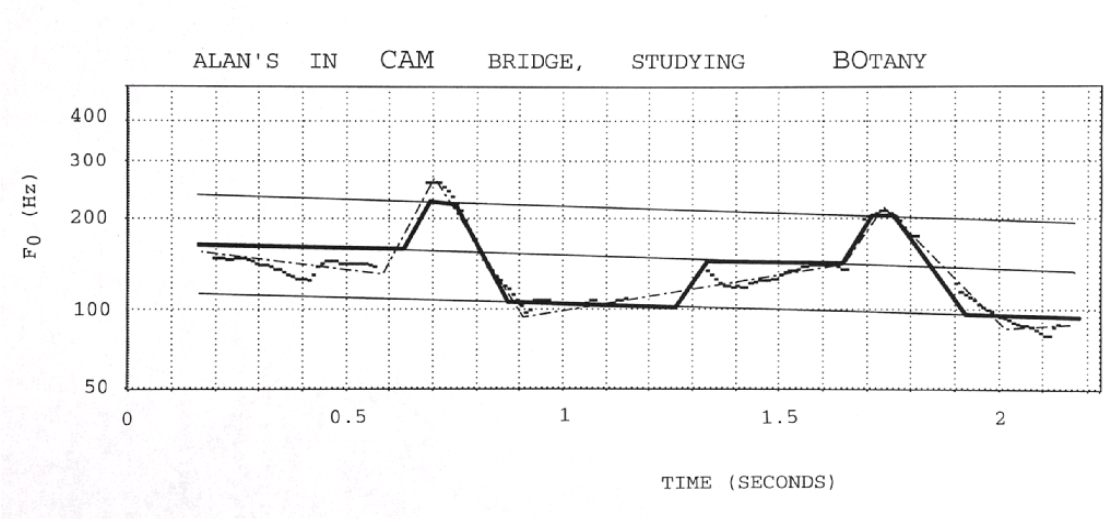

Fig. 3: Measured course of pitch (dotted line), close-copy stylization (interrupted line), a grid of three declination lines (solid lines), and standardized stylization (bold line) in a British-English utterance.

lines are slightly tilted, thus simulating the phenomenon of *declination*, making the pitch slowly drift down during all stretches where there are no abrupt changes in pitch that affect the perceived melodic properties of the utterance.

Declination has been attested as a rather general phenomenon in many languages. In its uninterrupted form it is generally constrained to rather brief sentences. In longer sentences, particularly in spontaneous speech, it is regularly interrupted by a “declination reset”, giving a new high start to the pitch of the next chunk of speech. It is assume by “t Hart et al. (1990) that the production of declination is not under

voluntary control, whereas intonation resets are voluntarily controlled by speakers. Other researchers are inclined to believe that at least part of the declination is due to a voluntary down step (Cf. Liberman and Pierrehumbert, 1984; Ladd, 1988). Speakers prefer to let these resets coincide with important boundaries in the constituent structure of the utterance (‘t Hart et al., 1990:Chapter 5). It is often not easy to estimate declination from pitch curves measured from naturally spoken utterances, because of the capriciousness of the pitch fluctuations, the variability of the size of pitch movements, and the occurrences of declination resets. ‘t Hart et al. suggest that the declination should be estimated by first making a close-copy stylization and then by trying to replace the stretches of relatively low pitch by pieces of a single straight-line, a tentative baseline. The tilt of this baseline would then give an estimate of the declination and would, in resynthesis, give the same perceptual result as the original declination.

The three equidistant lines in Fig. 3, to be called basic pitch levels (topline, midline, baseline) from now on, are typical for the description of British English (See also Brown et al., 1980). For other languages this may be different. For the description of Dutch intonation for example, it appears that two such basic pitch levels, a topline and a baseline, suffice as reference lines for defining virtually all perceptually relevant rises and falls (‘t Hart et al., 1990:76). Adriaens (1991) needed three equidistant basic pitch levels with one extra pitch line between the topline and midline, for defining the major standard pitch movements in German. Because pitch levels are equidistant in the time - log F_0 domain, distances can be fixed in semitones. For British English a distance of 12 semitones between baseline and topline gives satisfactory results. For Dutch, 6 semitones and for German 7.5 semitones between baseline and topline are adequate. One should notice that fixing these differences is a very severe reduction in the variability of pitch fluctuations. It is imaginable that for different styles of speech one should use different distances between pitch levels. Such changes would not, however, change the perceptual equivalence in terms of recognizable melodic patterns of the language.

Standard pitch movements can now be defined as more or less rapid transitions from one pitch level to another. Each standard pitch movement is fully characterized by its *direction* (up or down), its *size* (the number of semitones covered by the pitch movement), its rate of change (n semitones per second), and its timing (n ms after syllable onset or before syllable offset). Number and characterizations of standardized perceptually relevant pitch movements differ from language to language. For Dutch, ten perceptually relevant pitch movements are distinguished, for British English as many as 27 (Willems, ‘t Hart and Collier, 1988), and for German 11 (Adriaens, 1991).

Although the number of standard perceptually relevant pitch movements may differ from language to language, this number is always limited by the maximum number of categories that can be kept apart on each dimension. For direction this number is obviously not more than two, up or down. With respect to size it has been estimated that no more than three or four distinguishable intervals can be kept apart (‘t Hart, 1981). Rate of change allows, within the bounds of a single syllable, only for one rise (and presumably one fall) to be discriminated from nonchanging pitch. Only gradual rises that extend over several syllables may perceptibly differ in rate of change (‘t Hart, 1976). With regard to timing within the syllable, it appears that within a syllable of 200 ms, no more than three distinctive positions can be kept apart. The maximum number of abrupt pitch movements, occurring within a single syllable, for a particular language thus appears to be in the order of 2 x 4 x 1 x 3 = 24. This must be a comforting thought for students of intonation. Of course, the maximum number of possible pitch movements increases somewhat if pitch movements extending over more syllables (gradual rises and gradual falls) are included.

Perceptually, not all pitch movements have the same effect. Some rises and falls, such as the early rise and late fall in Dutch, serve as phonetic realizations of (pitch) accents, by lending perceptual prominence to particular syllables. Others, such as the late rise and early fall in Dutch, seem more suited to mark some kind of non-finality, either in mid-utterance or at the end of an utterance.

2.5 Combining pitch movements: towards a grammar of intonation.

Once one has defined an inventory of pitch movements for a particular language, it should be possible to generate sequences of such pitch movements. Such sequences would then constitute acceptable melodic realizations for speech utterances. Trying to do that, one will soon find out that not all possible sequences are acceptable. So, for example, in Dutch an accent-representing rise cannot be followed by another accent-representing rise without an intermediate fall. Also, studying the distribution of pitch movements in a corpus of utterances, it may become obvious that some pitch movements belong closer together than others: there appears to be a multilevel hierarchical structure to intonation. If we consider pitch movements themselves to constitute the lowest or first level of description, the second level is that of configurations, and the third that of contours.

A configuration is a close-knit intonational unit consisting of one or more consecutive pitch movements, for example a rise followed by a fall or a rise followed by a fall followed by a rise. Generally, constraints on combining pitch movements are much stricter within a configuration than at its boundaries. In their description of

Dutch intonation ‘t Hart et al. distinguish Prefix configurations, Root configurations, and Suffix configurations, notions that closely resemble the time-honoured notions of Head, Nucleus and Tail in the British impressionistic tradition of intonation studies. Prefix configurations are optional and recursive. They always precede another Prefix or a Root. Root configurations are obligatory and non-recursive: each contour must contain only one and not more than one Root. Suffix configurations are optional and non-recursive. A Suffix always follows a Root.

Pitch contours are defined as lawful sequences of configurations. Each pitch contour extends over a clause (in some loose sense of the word clause, referring to a group of words that the speaker has chosen as belonging together, as having some kind of coherence). This entails that multi-clause sentences have as many contours as there are clauses. Because there are recursive elements in contour formation, the number of contours is unlimited.

Many sequences of Prefix, Root and Suffix appear to be unlawful. Therefore explicit rules are needed to generate the lawful sequences and exclude the unlawful ones. The inventory of pitch movements, their combinations in configurations, plus the set of rules generating the lawful contours, together constitute a grammar of intonation. Ideally, such a grammar of intonation generates all and only the acceptable pitch contours of the language. The predictions by the grammar of Dutch intonation were verified against a corpus of 1500 spontaneous and semi-spontaneous utterances, and found to account for 94 percent of the contours in the corpus (‘t Hart and Collier, 1975).

2. 6 Basic intonation patterns

For both British English (Gussenhoven, 1983; Gussenhoven, 1984; Willems et al., 1988) and for Dutch (Collier and ‘t Hart, 1972; Collier, 1975) it has been shown that pitch contours can be classified into different families. Pitch contours belonging to the same family are put in the same class by native listeners when these are asked to sort utterances into a limited arbitrary number of subjective melodic categories. For both Dutch and English it appears that class membership is determined by one or more pitch movements belonging to the Root configuration. The pitch contours belonging to the same family are supposed to be manifestations of the same underlying “basic intonation pattern”. In the grammar of intonation each basic intonation pattern can be defined as the family of generation paths that go through the Root configurations that corresponds to that pattern. For both British English and Dutch six such basic intonation patterns can be distinguished, probably carrying different attitudinal and/or emotional connotations. For Russian some ten such basic intonation patterns have

been distinguished (Odé, 1986). Most of these basic intonation patterns are used rather infrequently. In Dutch, over 60 % of pitch contour tokens one encounters in everyday speech are realizations of a single basic intonation pattern, the so-called “hat pattern”.

2.7 Text and tune

So far intonation has been dealt with here virtually without reference to the sequences of words on which it is superimposed in actual utterances. In order to select a fitting pitch contour for a particular sentence, or sequence of pitch contours in the case of longer sentences, one has at least to know two things about the sentence. One has to know which words are to be provided with a pitch accent on their lexically stressed syllable, and whether, and if so where, boundaries between successive clauses in the sentences are to be made. Once these things are known, acceptable pitch contours can be selected for all clauses from the ones generated by the grammar of intonation. Of course, for each clause or sentence with known accent placements, there still is a variety of different possible pitch contours, and each pitch contour, due to its inherent flexibility with respect to time, can be made to fit a variety of different clauses or sentences.

In normal human speech the speaker determines which words are to be accented and where clause boundaries are to be made according to rules and strategies that will be briefly discussed in section 4 of this chapter. In synthetic speech, for example in text-to-speech systems, such rules and strategies have to be approximated by automatic text analysis (Kulas and Rühl, 1985; Carlson and Granström, 1986; Allen, Hunnicutt, and Klatt, 1987; Quené and Kager, 1993; Dirksen and Quené, 1993). Once this is done, appropriate and acceptably sounding pitch contours can be generated automatically and synchronized with the synthetic speech. Generally, in synthetic speech rules for generating pitch contours are limited to pitch contours that are manifestations of a single, neutral sounding, basic intonation pattern, as there is no basis to select between different intonation patterns.

The approach to the description of intonation sketched here, has the great advantage that it may lead to a set of rules that generate melodical equivalents to the vast majority of naturally occurring pitch curves in a particular language. This is achieved by severe reduction of reality by stylization and standardization. The result is that rule-generated equivalents of natural pitch curves, although on their own perfectly acceptable, are often much less lively than their natural counterparts. A long text read out with only synthetic pitch contours as generated by the grammar, may sound

somewhat dull and monotonous. Future research in this area might be directed at capturing generalizations that would reintroduce some of the natural liveliness in synthetic pitch contours, for example by varying excursion size and tone register.

3. THE RHYTHM OF SPEECH

3.1 Introduction

This section is concerned with the rhythm of speech. The very notion "rhythm of speech" suggests that two different utterances may share a common, underlying, property, called the same "rhythm". Intuitively, this can be brought to awareness by imitating the rhythmical pattern of an utterance with nonsense syllables, as "The MAN in the STREET" (where capitalized words are accented), imitated with "daDAdadaDA". Notice that one can do this at least in two different ways, either preserving the speech melody of the original utterance, or in a monotone. In case of the monotonous version we still can judge whether or not the imitation of the original rhythmical structure is successful. This suggests that it is possible, at least in first approximation, to study the rhythm of speech as a function of the temporal patterning of speech, without taking into account the melodic aspects.

As in the case of intonation, we will approach the rhythm of speech from the phonetic angle, concentrating on the ensemble of speech sound durations, that together constitute the temporal patterning of speech, attempting to focus on those aspects of temporal patterns that are relevant to the perceived rhythmical structure of speech, and de-emphasizing those aspects that are not. However, as will be shown below, the state of affairs with respect to rhythm is very different from the one in intonation. It will be made clear that different factors contributing to durational variation cannot so easily be separated. The primacy of perceptual structure, so helpful in the description of intonation to achieve a severe and useful reduction of capricious reality, teaches us that not only in production but also in perception temporal patterning is inherently complex and much less easily modeled. The corollary of this is that below we will not give primacy to a single unified approach to the study of temporal patterning, but rather will be more eclectic, asking attention for different approaches that may be seen as complementary. More particularly, we will discuss well-controlled experiments on so-called "reiterant" speech in 3.3, focussing on rhythmical speech patterns as it where in vitro, similar controlled experiments on real speech in 3.4, the many functions and quantitative interactions in temporal patterning in 3.5, statistical database studies, that seem to be indispensable as heuristic tools in this area, in 3.6, and rule systems for temporal patterning providing us with a perceptual testing ground in 3.7. But first we will discuss some basic aspects of the production and perception of speech sound durations in 3.2.

3.2. Speech sound durations: production and perception

Most sounds in nature are not indefinitely prolonged. They have an onset and an offset, and physical duration is determined by the time interval elapsing between onset and offset. Of course, perceived duration is determined by perceived onset and perceived offset, and an appropriate perceptual measure of time elapsed. This is not only so for the duration of sounds, but also for the duration of silent intervals between sounds.

In speech we rarely encounter isolated sounds. As speech develops in time, more or less abrupt changes in amplitude and spectral properties alternate with more or less homogeneous segments. Abrupt changes in the physical signal are caused by changes in the configuration of the vocal organs, such as opening and closing of the aperture of the vocal tract and onsets and offsets of vocal cord vibrations. Such changes demarcate both filled intervals, such as manifestations of vowels and fricatives, and silent intervals, as in manifestations of stop consonants. In oscillographic and spectrographic registrations of speech, where time is represented by spatial distance, one can measure the physical durations of such intervals. Fig. 4 gives an oscillographic representation of an English utterance “the queen said, the knight is a monster”.

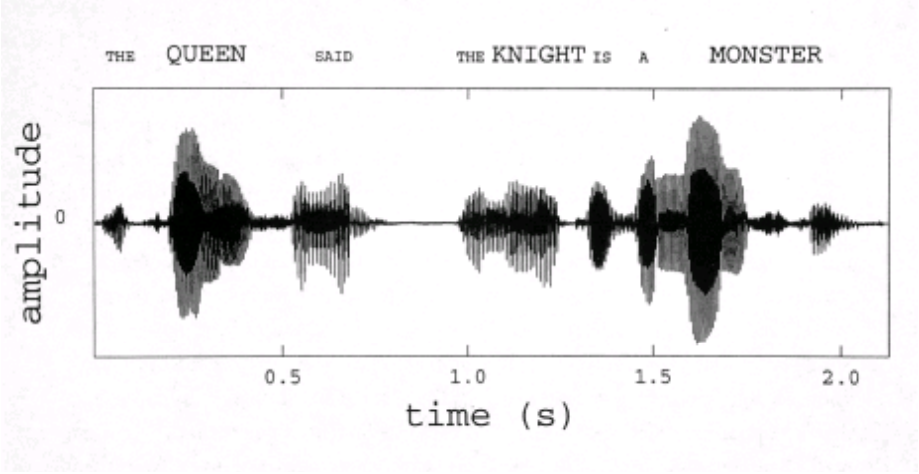

Fig. 4: Oscillographic representation of the utterance “The queen said, the knight is a monster”.

In such registrations we can measure physical durations of vowel-like segments, nasal segments, fricative segments, silent intervals, noise bursts, etc., by making some assumptions about relevant criteria for segmenting the signal. In case of clear and rapid changes in the signal such criteria are mostly straightforward, and segmentation accuracy is often more precise than human perception. In case of less clear, slower changes such as in glides or in slurred speech, there may be more uncertainty both for the investigator and to human speech perception. Notice that the perceptual relevance

of measured durations depends on the level of perception one is interested in. For example, the duration of formant transitions at the beginning of the vowel in a CV combination, being part of connected speech, may be relevant to the perceived consonant, but irrelevant to the contribution of the vowel-like segment as a whole to the perceived rhythm of speech.

Differential thresholds for isolated non-speech or speech sounds with durations in between approximately 40 and 250 ms are in the order of 5 to 15% (Ruhm, Mencke, Milburn, Cooper and Rose, 1966; Abel, 1972a), depending on the type of experiment. For silent intervals differential thresholds are somewhat higher (Abel, 1972b; Fujisaki, Nakamura and Imoto, 1973). Also, durations of silent intervals tend to be somewhat underestimated in comparison to those of filled intervals (Burghardt, 1973a; Burghardt, 1973b). Sound and silent interval durations shorter than about 40 ms and longer than about 250 ms are less accurately perceived than those in between these values. Very short intervals, shorter than roughly 40 ms for silent intervals and shorter than roughly 20 ms for filled intervals, do not seem to have subjective durations at all.

There is, mostly for practical reasons, relatively little research on just noticeable differences of speech sound durations being part of connected speech. However, it has been argued (Nooteboom and Doodeman, 1980) that we may infer the differential threshold, for example for vowel duration, from a binary forced choice phoneme classification task. In the experiment concerned the vowel of a Dutch word “taak”, embedded in a longer utterance, was given a number of different durations. Long durations led to perceiving “taak” with long /a:/, short durations to perceiving the word “tak” with short /A/. The transition of /A/ to /a:/ as a function of vowel duration could be modeled by a cumulative normal distribution with a mean (the phoneme boundary) of approximately 90 ms, and a standard deviation in the order of 5 ms. This suggests that the duration of embedded segments can, if the need arises, be perceived with an accuracy that is at least as good as that found for isolated sounds. Of course in an experiment like this listeners hear the same utterance over and over again, and may thus establish a fixed temporal reference pattern that may hone their ability to hear small differences (cf. O’Shaughnessy, 1987:160-161). Using a category judgment technique with nine durational categories for an embedded vowel and an embedded fricative, being part of a longer sentence, Klatt and Cooper (1975) estimated differential thresholds of 25 ms and more. The task is difficult, however, and may overestimate differential thresholds. Huggins (1972) found differential thresholds of even 40 ms, but it may be argued that the task he used, asking in an “up-down” strategy whether a particular duration in an utterance is longer or shorter than normal, measured perceptual tolerance rather than perceptual acuity. Of course, perceptual

tolerances may be more relevant to the purpose of this chapter than differential thresholds. One should notice, however, that perceptual tolerances measured for specific segments in specific utterances cannot easily be generalized to other segments and other contexts. Perceptual tolerances seem to vary considerably from one segment to another and one context to another in connected speech, and as yet we have no way of making adequate predictions of perceptual tolerances.

3.3 Prosodic temporal patterns: evidence from reiterant speech

Speech is not rhythmical in the strict sense that dance music is, with such a regular alternation of strong and weak elements in the stream of sound that the upcoming elements can be fairly precisely anticipated in psychological time from the preceding ones. Speech is rhythmical, however, in the looser sense that its development in time is controlled by some hierarchical mental pattern giving each syllable a certain strength that controls aspects of its production, among which is its duration. The resulting patterns are recognizable and can be imitated by users of the language, and lend speech an organization that helps its mental processing by the listener. Rhythmical imitations of speech utterances in sequences of identical nonsense syllables have been named "reiterant speech" by Nakatani and Schaffer (1978), but their use in speech research is much older than the name (Cf. Lindblom, 1968; Lindblom and Rapp, 1973; Nooteboom, 1972).

Fig. 5 shows temporal patterns obtained with reiterant speech for Dutch words spoken in isolation, varying from one to four syllables, with stress on the first syllable that contains either the long vowel /a:/ or the short vowel /A/. These data, among other things, exemplify the well known phenomenon of *compensatory shortening*: the more syllables follow the lexically stressed syllable within the same unit, the shorter its duration (and the durations of its segments). The data also clearly exemplify the phenomenon of *final lengthening*: the segment durations of the unstressed last syllable are considerably longer than those of other unstressed syllables. Very similar patterns, but somewhat shortened, are obtained for words embedded with longer utterances, spoken either with or without a pitch accent on the lexically stressed syllable (Nooteboom, 1972).

These and similar data clearly reveal a number of regularities, that cannot be attributed to the segmental make up of the syllables, being identical for each syllable, but apparently result from underlying mental patterns. The existence of such patterns is in accordance with the so-called "rhythm hypothesis" of Kozhevnikov and Chistovich (1965). In these patterns the word is a major unit. Lexically stressed syllables, whether accented or not, are considerably longer than non-final unstressed

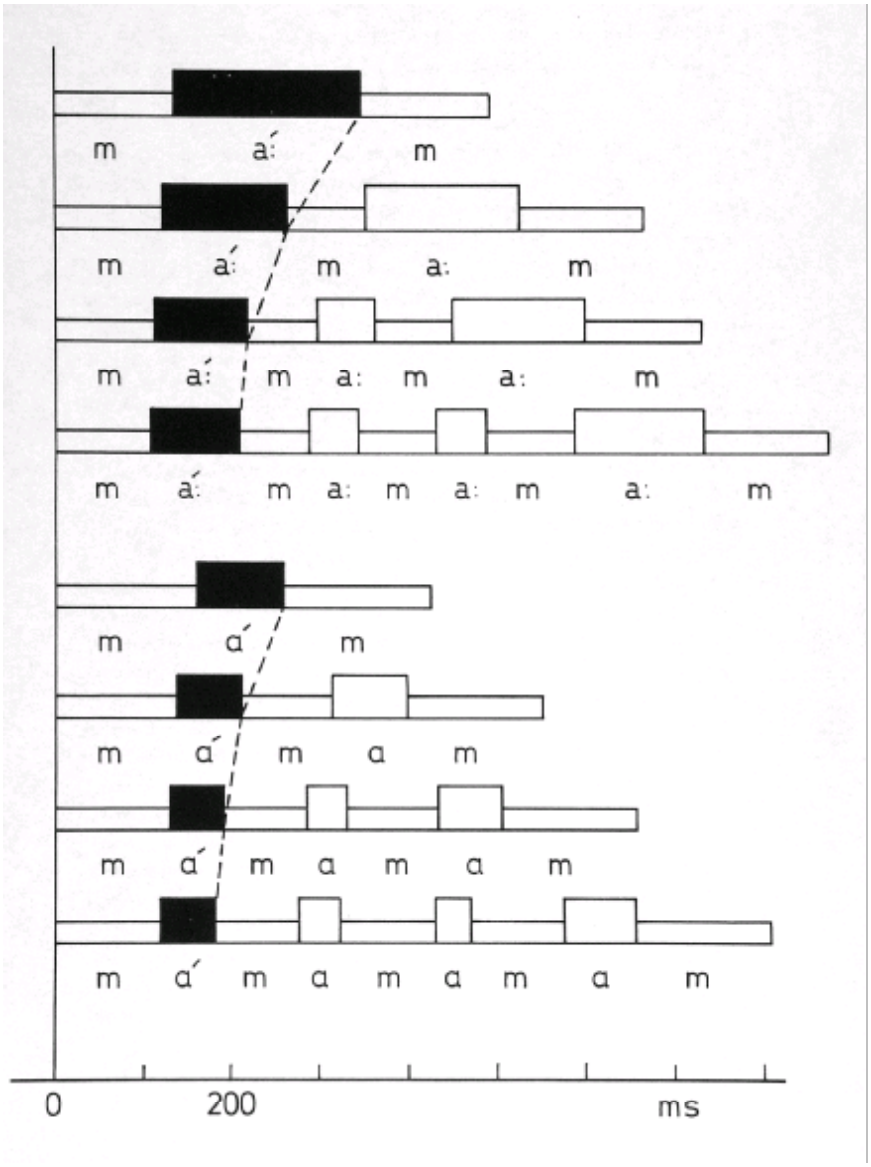

Fig. 5: Schematized temporal patterns of reiterant versions of Dutch spoken words with stress on the first syllable, and varying from one to four syllables. Top: with repetitions of the syllable [ma:m], bottom: with repetitions of the syllable [mɑm].

ones and exhibit considerable *anticipatory compensatory shortening* (as in Fig. 5) and some *perseveratory compensatory shortening* (the more syllables precede the stressed syllable, the shorter the stressed syllable becomes). Other things being equal, accented lexically stressed syllables are somewhat longer than unaccented ones. The word-initial unstressed syllable is somewhat longer and the word final syllable considerably longer than the medial unstressed syllables. The pattern appears to be cyclical: it is repeated on the level of phrases, accented words being longer than unaccented ones, phrase initial words being somewhat longer and phrase final words considerably longer than phrase medial words (Lindblom and Rapp, 1973). Partial evidence for perceptual relevance of such patterns has been obtained by Nakatani and Schaffer

(1978), who demonstrated that word boundaries in trisyllabic adjective noun combinations can be perceived from reiterant versions of such short phrases. De Rooij (1979; see also Nooteboom, Brokx and De Rooij, 1978) demonstrated that under certain conditions constituent boundaries can be adequately perceived from reiterant speech on the basis of temporal cues. The latter appeared to be much more effective in this respect than melodic cues.

3.4 Confirmation from real speech

Experiments with reiterant speech are nice because they are revealing of regular underlying mental patterns that otherwise would remain obscure due to the extreme variability of speech sound durations in connected speech. However, the regularities obtained should be interpreted in a qualitative rather than in a quantitative sense. Whether and how these regularities quantitatively show up in real speech depends on many factors, as will be argued in 3.6 below. A first reassurance we need is whether similar temporal regularities can be demonstrated for real words and phrases, and whether such regularities can be shown to be part of what language users (implicitly) know about the way words and phrases in their language should sound. Both issues are, by way of example, addressed in Fig. 6 with respect to the phenomenon of compensatory shortening.

The figure plots durations of the vowels, long /a:/ or short /A/, of lexically stressed initial syllables of real Dutch words varying from one to four syllables, as a function of the number of syllables in the word. Three types of durations are plotted: durations measured in these words spoken in isolation, durations obtained in a method of adjustment with synthetically spoken versions of these words (see below), and durations calculated from a simple empirical rule $V = D/m^{\alpha}$, in which V is the vowel duration to be calculated, D is a standard duration for the vowel concerned, m is the number of syllables following in the word plus 1, and α is a constant with the value 0.2 for both long /a:/ and short /A/. Clearly compensatory shortening is strongly present in the spoken versions and well described by the empirical rule. In the method of adjustment used to obtain the third kind of data, subjects were asked to adjust, by turning a knob, the durations of the stressed vowels in synthetic versions of these words such that the word as a whole sounded as natural as possible. Again the pattern of compensatory shortening is accurately reproduced, showing that such regularities are part of the mental representation of speech. Similar results were obtained for a number of the regularities discussed in 3.3, such as stressed versus unstressed and initial and final lengthening (Nooteboom, 1972 and 1973), demonstrating that, at least

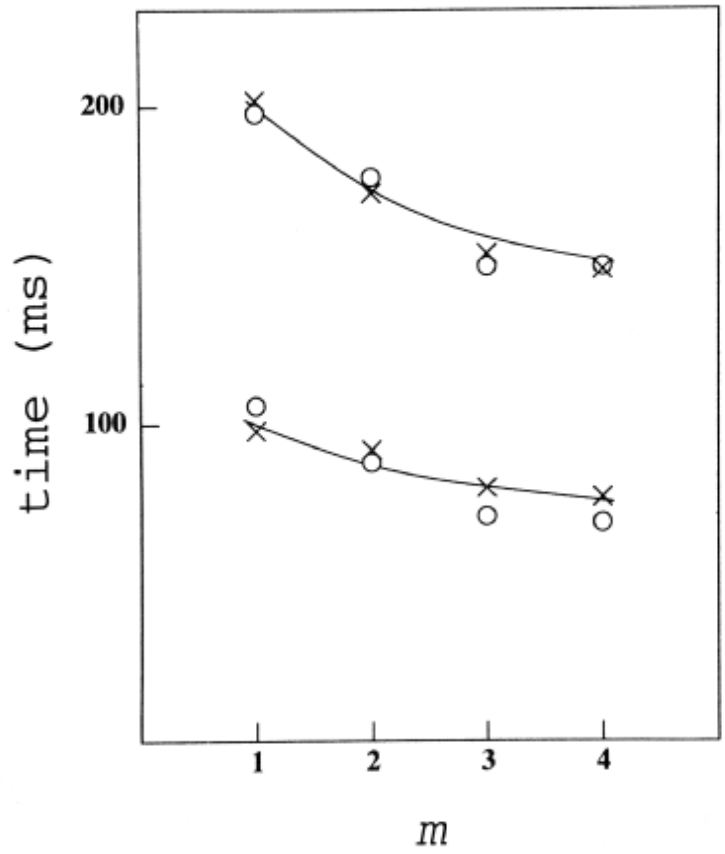

Fig. 6: Calculated (solid line), spoken (crosses) and adjusted (circles) durations of stressed [a:] (top curve) and [ð] (bottom curve) as a function of the number of syllables in the word which remain to be produced at the beginning of the syllable concerned.

under certain conditions, patterns found in reiterant speech also show up in real words and that they are psychologically real.

3.5 Multifunctionality and interactions in the temporal patterning in speech

In the study of intonation we have seen that the natural course of pitch can be supplanted with a highly stylized approximation without noticeable changes to perception. This finding formed the foundation of highly simplified and standardized melodic models of the perceptually relevant structure of intonation. These models can immediately be used in synthetic speech virtually without adaptation to the segmental structure of speech, except for some simple synchronization rules. It would be nice and easy if the same trick applied to speech rhythm. We can try this out by

manipulating a naturally spoken speech utterance of some complexity as follows. We have someone speak a reiterant version of that utterance. Both original utterance and the reiterant version are analyzed by means of Linear Prediction (see Chapter XX of this handbook). In both original and reiterant version we mark each syllable boundary. After that we shrink and stretch, before resynthesis, all syllable durations in the original utterance so that they obtain the durations of the corresponding syllables in the reiterant utterance (Fig. 7). The result of this exercise in many cases is disastrous. Not only are the differences between the two versions of the same utterance (top and bottom in Fig. 7) easily perceived, but the perceived rhythmic pattern changes completely, and may become highly unnatural. The reader may note that in this demonstration experiment it is implicitly assumed that syllables are the basic units of timing in speech. It has also been suggested that rather intervals from vowel onset to vowel onset determine the perceived rhythmical patterns in speech (Huggins, 1968). Redoing the experiment lining up intervals between successive vowel onsets will produce results that are, although not identical, very similar. Either way, this simple class room experiment immediately confronts us with the relative inviolability of temporal patterning of speech.

The reason for our failure is to be found in the multifunctional nature of speech sound durations, that are affected by a great many very divergent factors in production, and affect a great many very divergent perceived aspects of speech. Speech sound durations are affected by and carry information on both within syllable factors and between syllable factors. Examples of within syllable factors are segment identity, and the identities of preceding and following phonemic segments. Segment identity is, in many languages, involved in the opposition between phonologically long and short phonemes. On a more phonetic level, we observe for example that open vowels like /a/ have, other things being equal, longer durations than closed vowels like /e/ or /o/, simply because it takes more time to open the mouth further than to open it less (House, 1961; Nooteboom, 1972). It has also been found that the closed interval of a voiceless stop is much longer than the one in a voiced stop, and that this differences contributes to identification (Lisker, 1957; Slis and Cohen, 1969a and 1969b). A clear example of an effect of the following phonemic segment is that the vowel preceding a voiceless stop consonant is longer than the one preceding a voiced stop within the same syllable. This difference contributes to the perception of stop voicedness (Lisker, 1957; Slis and Cohen, 1969a and 1969b). But we also see that the duration of the silent interval of a stop consonant is affected by the preceding vowel segment: after a long vowel this duration is markedly shorter than after a short vowel. In perception the duration of a closed interval of a stop consonant can affect the

perceived phonological length of the preceding vowel (Nooteboom, Brokx and De Rooij, 1978).

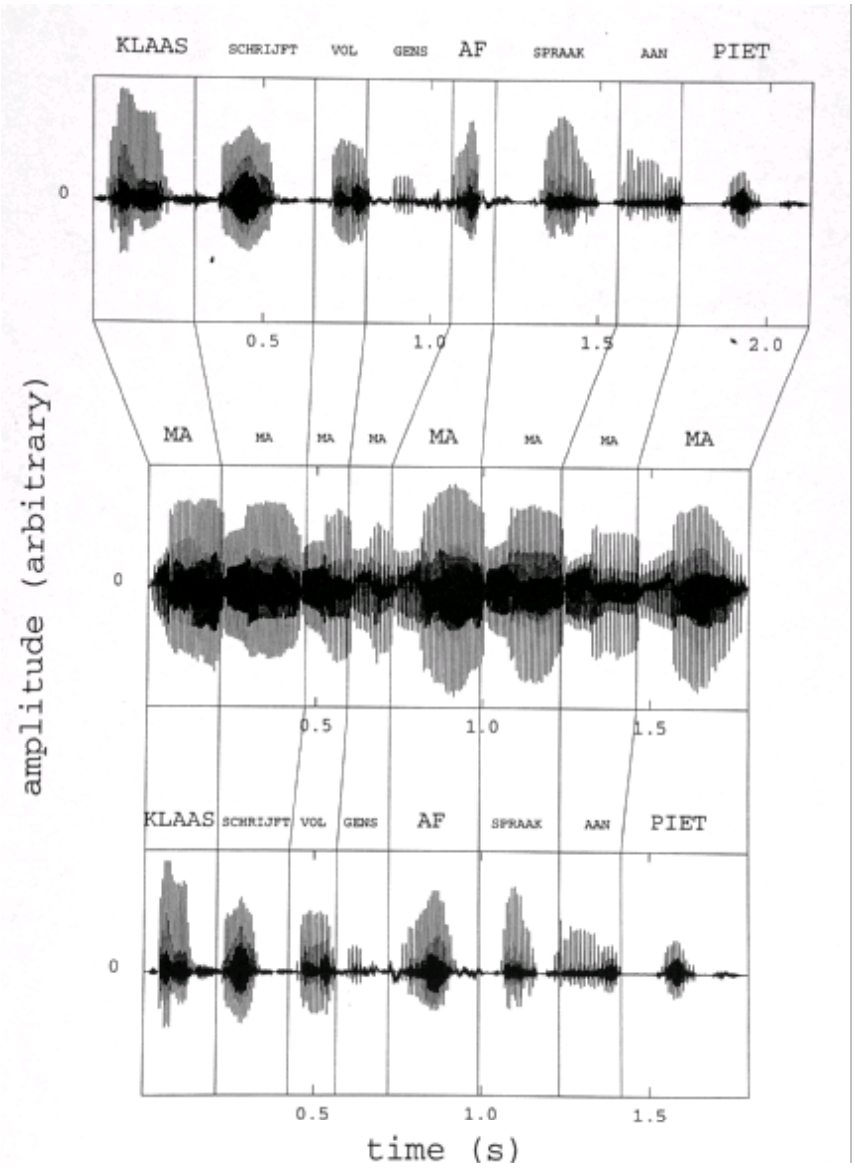

Fig. 7: Three oscillograms. Top: an originally spoken Dutch utterance. Bottom: a reiterant version of the same utterance, each syllable being spoken as [ma:]. Bottom: Same as top, but with syllable durations made identical to those in the reiterant version.

Examples of between syllable factors are the relation between overall vocal effort and segment durations, the effects of lexical stress and accent, the effects of sentence, phrase and word boundaries, and the effect of rhythmical alternation in sequences of of unstressed syllables. It has been shown that as vocal effort increases, vowel durations increase and consonant durations decrease. These differences are related to the wider opening of the mouth in loud speech compared to normal speech. It

appeared impossible, however, to model these differences in a linear scaling model of the behaviour of lips and jaw. This suggests that there is extensive reorganization of articulatory behaviour, by which other perceptually relevant aspects of temporal patterning remain better preserved (Lindblom, 1989). Other things being equal, lexically stressed syllables are often considerably longer than lexically unstressed syllables, although this difference itself depends much on position within word and phrase. Perception of lexical stress depends to a large extent on the pattern of syllable durations (Lindblom, 1968; Nooteboom, 1972; Nakatani and Schaffer, 1978). Over and above the effect of lexical stress there is a considerable effect of sentence accents or phrasal accents. Eefting found that in prepared speech accented words are roughly 20 % longer than unaccented ones. This difference appears to be equally distributed over lexically stressed and unstressed syllables. It cannot be tampered with without reducing the perceived acceptability of speech and speed of processing (Eefting, 1991). Segments at word boundaries tend to be somewhat longer than segments within words. This difference contributes to word boundary detection (Nakatani and Schaffer, 1978; Quené, 1992; Eefting, 1991). Quené (1989) showed that in spoken versions of word pairs such as “known ocean” versus “no notion”, excised from the utterances they were spoken in, word boundaries can be detected with 80 % accuracy on the basis of temporal patterning.

Segments in syllables immediately preceding sentence boundaries and major and minor phrase boundaries in the stream of speech are considerably longer than segments in other syllables, other things being equal (Klatt, 1976; Nooteboom, Brokx and De Rooij, 1978; Van Santen, 1992; Campbell, 1992). Slootweg demonstrated that in sequences of three unstressed non-word initial and non-word final syllables in one spoken word, the middle one is rhythmically stronger and has a longer duration than the other two (Slootweg, 1988). Lehiste (1970) points out that, at least in some languages, speech tempo affects syllable durations differentially. In English, unstressed syllables are affected more than stressed syllables. It has been assumed that this is typical for so-called stress-timed languages in contrast to so-called syllable-timed languages (Bertinetto, 1981). Stress-timed languages are supposed to have a tendency to make intervals consisting of a stressed syllables plus all following unstressed one equally long; syllable-timed languages are supposed to have a tendency to make all syllables equally long. Den Os (1988) found little reason for this distinction in a systematic comparison between spoken Dutch, an alleged stress-timed language, and spoken Italian, supposedly a syllable-timed language. Presumably, the intuitive difference between stress- and syllable-timed languages has its origin in the presence versus absence of a phonological length opposition, and/or the presence

versus absence of vowel reduction, and other aspects of the segmental make-up of syllables, such as the number of consonants permitted in consonant clusters.

The fact that there are so many different factors acting simultaneously on the perceptually relevant temporal patterns of speech makes life hard for speech researchers who want to give a systematic account of such temporal patterning. Life becomes even harder because there appear to be strong quantitative interactions between various factors. The well-known effect of compensatory shortening (see 3.3 and 3.4) on vowel durations is very different, both in milliseconds and in percentage, for different vowel phonemes, depending on their "compressibility". Not only vowel identity, but also the identities of preceding and following consonants affect vowel compressibility (Klatt, 1976), thereby changing the quantitative effect of compensatory shortening. This quantitative effect changes again, both in absolute and in relative terms, when speech tempo is changed. The higher the speech tempo, the less shortening we find (Nooteboom, 1992). Very brief vowels, like /I/ before voiced stops, may not show any appreciable compensatory shortening at all, although in that case the syllable as a whole may still be shortened somewhat by shortening consonant durations surrounding the vowel.

The durational difference between lexically stressed and unstressed syllables and their segments is also far from fixed, and depends both on the type of segments in the syllable and on syllable position in word and utterance. In word medial and utterance medial position this difference between stressed and unstressed syllables may be considerable, in the order of hundreds of milliseconds, but in prepausal position the difference often is negligible, apparently because prepausal lengthening has exhausted the "stretchability" of the syllable. It is as if each particular segment type within a particular syllabic environment can only vary its duration between a maximum and a minimum that are typical for that segment in that environment (Klatt, 1976).

Such extreme interactions between many different factors affecting syllable and segment durations (of which many more examples are known, cf. Van Santen, 1992) have the result that the systematic effects on speech sound durations of any one particular factor can only reliably be assessed when we take the effects of many other factors into account (Nooteboom, 1992). This has some consequences. One is that we cannot study or model the quantitative effects of rhythmical factors, presumably coinciding with the "between syllable factors" listed above, in isolation from "within syllable factors". Another is that one needs rather large databases and statistical tools to get an appropriate impression of the factors involved and their quantitative interactions. One also needs sophisticated quantitative models to account for the effects of these factors and their interactions.

So far we have concentrated on speech sound durations in stretches of speech between speech pauses. Speech pauses themselves are an integral part of temporal patterning, and play an important role in speech perception. Notice that one should make a distinction between acoustic silent intervals and perceived speech pauses. Not all silent intervals are perceived as speech pauses. For example silent intervals as part of the production of voiceless stop consonants are generally not perceived as speech pauses, unless their duration is abnormally long. Also subjectively speech pauses may be perceived where there is no silent interval, provided there is considerable final lengthening and/or a clear melodic boundary marker (Nooteboom, Brokx and De Rooij, 1978). So a subjectively perceived speech pause is triggered by final lengthening, whether or not it is accompanied by a melodic boundary marker (particular pitch movement and/or declination reset; see 2.4), and whether or not it is accompanied by a silent interval. Speech pauses, as cues to prosodic boundaries, are regularly used to demarcate major or minor phrases, and the particular acoustic realization seems to depend on the relation between the prosodic boundary and its position in the hierarchical constituent structure of the sentence being spoken (Harris and Umeda, 1974; Grosjean, Grosjean, and Lane, 1979; Cooper and Paccia-Cooper, 1980; ‘t Hart et al., 1990). Production of speech pauses is to a large extent optional, and depends much on style of speech and speech tempo (Goldman-Eisler, 1968).

There are also quantitative interactions between the duration of silent intervals as cues to speech pauses on the one hand and the production and perception of segment durations in the preceding syllable on the other. The longer the silent interval, the more final lengthening is produced by the speaker and expected by the listener. This expectation can affect the perception of phonological length of the syllable nucleus (Nooteboom and Doodeman, 1981). There is also another kind of interaction: the actual durations of silent intervals at boundary positions in longer spoken sentences appear to be predictable from the average stress group duration (time interval between vowel onsets) in the preceding stretch of speech plus the number of phonemes in the stress group containing the speech pause (Fant and Kruckenberg, 1989). Fant and Kruckenberg also found that silent interval durations at prosodic boundaries are not stochastically distributed but seem to cluster around certain values, the longer ones being multiples of the shortest one.

3.6 Quantitative approaches to the study of temporal patterning

Researchers have many reasons to study the temporal organisation of speech. These reasons may be of a fundamental scientific nature, or may be technologically oriented. The first, more fundamental, reasons include for example a desire to elucidate the

social, mental, physiological, or acoustic processes in speech communication (production and/or perception), or to find evidence for the psychological reality of linguistic units, patterns or rules. The second, more technological, reasons may stem from a wish to improve speech synthesis-by-rule or automatic speech recognition. This division between fundamentally motivated and technologically oriented research goals seems to be parallelled by a distinction between two types of data gathering. Fundamental research questions are generally approached with well controlled, but limited, stimulus materials, carefully designed for testing specific hypotheses. Examples abound in the literature. At the other end of the data-gathering dimension are statistical studies of speech segment durations based on more or less extensive corpuses of real, often connected, speech. Typically, such studies stem from technologically oriented research. Examples are provided by Barnwell (1971), Harris and Umeda (1974), Crystal and House (1982, 1988a, 1988b, 1988c, 1988d, 1990), Fant and Kruckenberg (1988a, 1988b, 1989), Fant, Nord and Kruckenberg (1986, 1987), Campbell (1990) and Van Santen (1992).

Van Santen basically used as a corpus a set of isolated sentences, together containing 13,048 word tokens, spoken by a single male speaker. He investigated quantitative effects on vowel durations of the following seven factors:

- Vowel identity
- Identities of the surrounding segments
- Position of the vowel in the syllable: left- and right-open versus closed syllables
- Position of the syllable within the word (number of syllables that precede and follow in the word); or within the stress interval (the number of unstressed syllables that precede and follow the target vowel in the sentence)
- Stress status of the syllable
- Position of the word in the sentence: effects of phrase boundaries
- Accent status of the word.

(The effects of speaking rate and syntactic structure were explicitly not analysed).

Except for stress intervals, all factors were found to have a considerable effect on vowel durations. Durations predicted from an eight parameter model incorporating these factors showed a correlation of 0.9 with the observed durations, accounting for 81% of the variance.

Such statistical studies of speech sound durations have shown a number of systematic regularities in the temporal organisation of speech, sometimes confirming, at other times seemingly contradicting earlier findings. For example, Crystal and House (1989) found no evidence for compensatory shortening in American English.

This is in agreement with the findings of Umeda (1975) for rapid connected speech, but is in contrast to what Harris and Umeda (1974) found for slower speech and in also in contrast to the findings of Van Santen (1992).

Statistical studies of speech sound durations also consistently confirm that there are strong quantitative interactions between different factors affecting speech sound durations. These interactions can be modelled by equations combining additive terms with multiplicative terms. A well known example is the empirical rule proposed by Klatt (1976), which in its simple form can be written as:

$$DUR = k(D_{inh} - D_{min}) + D_{min}$$

in which:

DUR is the segment duration to be calculated,

k is a parameter describing a context effext, or any combination of such parameters,

D_{inh} is a table value standing for the segment-specific inherent duration,

D_{min} is a table value standing for the segment-specific minimal duration.

In this model, context parameters provide a multiplicative term, and segment-specific parameters provide additive terms. Furthermore, context parameters are functionally combined, under the implicit assumption that the order of the joint effects of these parameters is unaffected by other factors. The model was until recently never rigorously tested. Van Santen and Olive (1989) show how to generalize models of this type mathematically and how such models can be tested by analyzing the covariances between sub-arrays of a multifactorial data matrix. Van Santen and Olive applied their method of model analysis to a data base containing 304 different phrases of two nonsense words, read by one male speaker at two speaking rates. They showed for vowel durations that, contrary to Klatt's model, in their database the segment-specific factors need only a multiplicative term, and the context factor both a multiplicative and an additive term. They also showed that no factors could be functionally combined.

This approach is interesting because it allows the researcher to tune both the mathematical form of the model and the values of its parameters to real data. If this method of analysis could be applied to large databases of real connected speech, it holds the promise that we may finally come to grips with the complex and until now obscuring interactions between many factors that affect speech sound durations. Evidently, this will not only be of advantage to speech synthesis-by-rule, or automatic speech recognition, but may also provide an interesting testing ground for predictions

made by models of speech production, and a rich source of testable hypotheses concerning speech perception.

Of course, there is no practical way in which we can be sure that the list of factors taken into account in this kind of modelling is exhaustive. Furthermore, the approach itself is not based on theories or models of speech production and speech perception, but rather represents the regularities in a database in the form of empirical rules. Predicting the effects of 'new' factors, that is factors that have hitherto been overlooked, should come from theoretical accounts of the mental, physiological, and acoustic processes in speech production and perception. Such accounts are often partial, not being part of any complete theory or model of all the processes involved in the production and perception of speech, and quantitative predictions are often very specific. Because of this, there remains a need for testing such predictions in vitro, in specifically designed laboratory experiments with well controlled stimulus materials. An approach as proposed by Van Santen and Olive can never replace, but rather is complementary to, theoretical and experimental studies of different processes in the production and perception of speech.

3.7 Speech rhythm and rule systems for the temporal patterning of speech

In terms of rule systems for the temporal patterning of speech that could be used for example in speech synthesis by rule, we can operationalize the rhythmic rules as those rules that take care of aspects of temporal patterning that, at least qualitatively, show up in reiterant versions of real utterances. The remaining rules could be named syllable production rules. We can then imagine a rule system that first takes care of syllable production, producing an idealized temporal pattern for each syllable to be spoken, and then applies the rules responsible for speech rhythm. In fact, there are quite a few rule systems that are organized this way (Campbell and Isard, 1991). One should realize, however, that due to the strong quantitative interactions discussed earlier, the rhythmic rules cannot be formulated in terms of constant additive or multiplicative values. Instead, they should contain parameters the values of which can only be filled in by going back to the level of syllable structure. Conversely, some aspects of syllable structure, such as assimilation, degemination, coarticulation and reduction, can only be adequately decided on after the rhythmic rules have applied, and the actual course of pitch generated in the intonation module can only be synchronized with the utterance after the complete temporal pattern has been generated. Generally the relations between parameter values and conditioning factors are given in lookup tables of considerable elaborateness, but in current rule systems

the two-way interactions are often neglected. Due to this, synthetic speech, although intelligible, sounds often unnaturally over-articulated, and pitch fluctuations are sometimes inadequately synchronized with the segmental structure of the utterance. There is considerable room here for further study of the interactions between different aspects of speech production and for modelling such interactions.

The difficulty in setting up adequate rule systems for the temporal patterns of speech derives from the fact that the sequence of speech sound durations code so many different things simultaneously and interactively. This obscures the relation between the rather simple abstract rhythmical patterns of speech on the one hand and their realizations in speech as it is produced on the other. As illustrated in the beginning of this section on speech rhythm, it is deceivingly simple for humans to recognize and imitate rhythmical patterns of speech utterances. To have a machine perform the same feat is as yet far beyond our capabilities.

4. SOME COMMUNICATIVE FUNCTIONS OF SPEECH PROSODY

4.1 Introduction

What is the use of speech prosody in normal speech communication? In normal written or printed text there is, apart from punctuation and the use of capitals, very little that corresponds to prosodic patterns in speech. Yet many people easily read more words per minute than speakers can speak at their fastest rate. There is, however, a major difference between text and speech. Text is spatially presented, such that much of it is simultaneously present to the reader. Speech is not. At each moment in time the sound of speech is nothing more than a momentary disturbance of air pressure. One moment it is there, the next moment it is gone. Because speech is often listened to in the presence of other sounds, continuously decisions have to be made which successive sounds are to be integrated in the utterance being perceived and which are to be rejected as extraneous. Here prosody helps (4.2).

The fleeting nature of the sound of speech also has the consequence that human perceptual processing of speech draws heavily on human short term memory functions. It is all in the mind. A listener cannot go back to the physical stimulus during processing, because that stimulus has forever vanished in the past. Yet we notice than in normal speech a great many phonemes are very rapidly produced, becoming grossly degraded to the extent that they become unidentifiable without context, or even are completely deleted. We may imagine that if this were not so, speech would become much too slow for the listeners to keep attention focussed on the contents of the message. As we learn from the comparison with reading, comprehension can go much faster than speech allows. But the less specified segmental structure is, the more support a listener needs from suprasegmental, prosodic cues. These cues can differentiate between more important and less important information as coded in accent patterns (4.3), and also organize the message in chunks that are easily processed by the listener, at the same time revealing aspects of the linguistic structure of the message (4.4).

These examples of communicative functions of speech prosody will be briefly described below. The list is not exhaustive. Prosody may to a certain extent also be used to characterize utterances as statements, questions, or exclamations (Hadding-Koch and Studdert-Kennedy, 1964; Geluykens, 1987; ‘t Hart et al. 1990, pp. 111ff), to convey information on attitude and emotion (Crystal, 1969; Murray and Arnott, 1993), or to characterize certain styles of speech.

4.2 Auditory continuity and the separation of simultaneous voices

Cherry (1953) addressed himself to the question of how one recognizes what one person is saying when others are speaking at the same time, a phenomenon he referred to as the 'cocktail party effect'. Cherry mentioned as possible facilitating factors directional hearing, visual information, individual differences in voice characteristics and dialect and transitional probabilities. Although his main experiments were directed at directional hearing and transitional probability, he also observed that, when all the above-mentioned factors except transitional probability were eliminated by recording two messages spoken by the same speaker on the same magnetic tape, the result may sound "like a babel", but the messages can still be separated.

Darwin (1975) neatly demonstrated that pitch continuity is an important factor in "voice tracking". He presented listeners simultaneously with two different passages of speech, spoken by the same speaker, either or not switching from one ear to the other and vice versa during presentation. The stimulus material was so constructed that four conditions were obtained, a normal condition with no switch, a semantic change condition, in which pitch was continuous on each ear but the verbal message switched ears in the middle, an intonation change condition where the verbal message was continuous on each ear, but intonation switched ears in the middle, and a semantics and intonation change condition, in which both verbal message and intonation switched ears simultaneously. Listeners were instructed to attend to one ear only. Switching the intonation from one ear to the other caused a high percentage of intrusions of the unattended ear, showing that listeners track a voice in the presence of another voice (and in the absence of directional cues) mainly on the basis of pitch continuity.

From Darwin's experiment it is reasonable to assume that perceptual separation of simultaneous speech messages is easier for messages in different pitch ranges than for messages in the same pitch range, where the listener may inadvertently switch to the other message whenever the two pitches cross. This was shown to be correct by Brokx and Nooteboom (1982), in an experiment with resynthesized speech utterances from a single speaker, with artificially manipulated pitches. There were approximately 20% less word perception errors with different than with the same pitch ranges.

Obviously, intelligibility of speech in the presence of other speech is better when the pitches or pitch ranges of the two competing messages are different than when they are the same. This effect can be related to the phenomenon of "perceptual fusion", occurring whenever two simultaneous sounds have identical pitches, and to "perceptual tracking": whenever the pitches of target and interfering speech cross

each other, the listener runs the risk of inadvertently switching his attention from the target to the interfering speech.

4.3 Accent patterns and their role in speech communication

In the act of speaking, some words are accented by means of an accent-representing pitch movement on their lexically stressed syllable, with some concomitant cues such as some extra loudness and some lengthening of the word. Of course we can establish that a particular word is accented without worrying too much how the accent is realised in the act of speaking: the notion 'accent' is abstract with respect to its realisation, and as such basically refers to the same thing as the notions 'sentence stress' or 'word group stress' (Chomsky and Halle, 1968; Liberman, 1979; Selkirk, 1984).

To show how accents are used in speech communication we need to introduce the notion 'focus', used here as in Ladd (1980), Gussenhoven (1983), Selkirk, (1984) and Baart (1987). A constituent, which can be a single word but also a word group or phrase, can be presented by the speaker as in focus (or +focus) by means of an accent on a single word that we call the prosodic head of the constituent. The position of the prosodic head within each potential constituent can be derived from syntactic structure. The reasons why a particular constituent can be put into focus, and thus receive an accent on its prosodic head, do not seem to be particularly well understood. But one of these reasons appears to be the 'newness' to the listener of the information contained in that constituent. Compare the following examples (accented words are capitalized):

(Who wrote that novel?)
(a) The dean of our FACULTY wrote that silly book

(Who wrote that novel?)
(b) The DEAN of our faculty wrote that silly book

In (a) the whole constituent "The dean of our FACULTY" is put into focus by the single accent on "FACULTY". the phrase "wrote that silly book" contains given or presupposed information and therefore stays out of focus. In (b) only "The DEAN" is brought into focus, because accentuation rules operating on the syntactic structure of constituents, determine that "DEAN" can never be the prosodic head of the whole constituent "the DEAN of our faculty". Obviously, the speaker presupposes that the listener knows that the author is someone of our faculty.

Speakers differ in the ways they use accent patterns, and often violate expectations of professional linguists in doing so. Still, it has been shown in acceptability experiments that in the ears of the listeners new information can hardly ever be acceptably associated with '-focus', whereas given information can rather often, although not always, acceptably be associated with both '-focus' and '+focus'. '+Focus' cannot only be perceived as signaling new information, but also as highlighting thematic relations with the context (Nooteboom and Kruyt, 1987). It has also been shown, in a speeded verification task in which listeners had to judge whether a particular utterance was or was not an accurate description of some situation on a visual display, that violation of the relation between newness versus givenness and accent patterns slows down comprehension (Terken and Nooteboom, 1987). Apparently, accent patterns play their modest part in speeding up human processing of speech.

4.4 Prosodic boundaries

A sequence of words like "the queen said the knight is a monster" can be read and spoken in at least two different ways: "the queen, said the knight, is a monster", or "the queen said, the knight is a monster". The ambiguity inherent in this sequence of words is disambiguated in speech by prosodic phrasing, producing either a strong prosodic boundary after "queen" and "knight" or after "said" (cf. Fig. 8).
Even when no such strong ambiguities are present, nevertheless speakers tend to divide their speech into prosodic phrases. Potential positions for prosodic phrase boundaries can be derived indirectly from syntactic trees, by assigning metrical trees to the syntactic trees, and then applying some simple phrasing rules to the metrical trees (Selkirk, 1984; Nespor and Vogel, 1986; Dirksen and Quené, 1993). Selkirk, and also Nespor and Vogel, distinguish between two types of phrases, I-phrases or intonational phrases, separated by major boundaries, and Phi-phrases or phonological phrases, separated by minor boundaries. Phi-phrases are combined into the hierarchically higher I-phrases. Dirksen and Quené, in the context of a text-to-speech system, attempt to implement some of the ideas of Selkirk and Nespor and Vogel in a set of computational rules. Instead of two hierarchically ordered types of phrases, they assume only one type of phrase boundary which may or may not be realized by a speech pause or final lengthening. Their phrasing rule, applied to metrical trees, states simply that a phrase boundary occurs between A and B if:

a. A and B are sisters
b. B is an XP, and
c. both A and B are accented,

where A and B are adjacent phrases, and XP is a maximal projection of an NP, VP or AP. The net result of this rule is that phrase boundaries are placed between 'major' phrases.

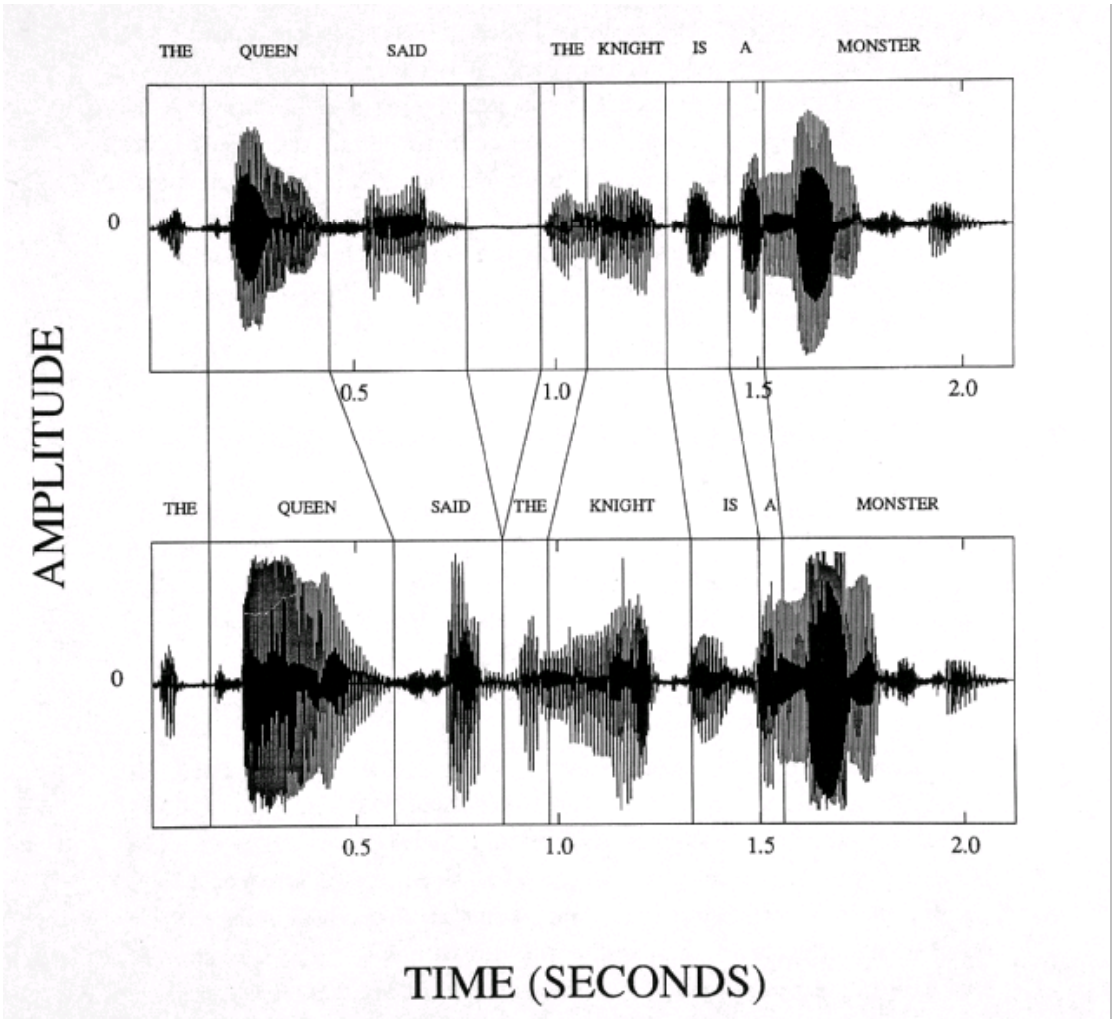

Fig. 8. Two oscillograms. Top: "The queen said, the knight is a monster". Bottom: "The queen, said the knight, is a monster".

Actual speakers assume considerable freedom in choosing whether phrase boundaries are or are not realized by speech pauses. They are more liable to make speech pauses in slow and careful speech than in rapid and less careful speech. Whenever they make a melodically marked speech pause, however, it is likely to be made at a predictable phrase boundary. Speech pauses that are only realized in the temporal pattern and not in the speech melody, appear to be less closely related to syntactic and metrical structure, and may either be hesitation pauses, or pauses used to other stylistic ends (Blaauw, unpublished).

High quality speech without grammatical speech pauses within sentences can be highly intelligible and acceptable. But as soon as speech quality is less than normal, or speech is listened to in noisy conditions, the introduction of grammatical speech pauses can help to maintain intelligibility (Nooteboom, 1985). In general, it can be observed that the contributions of prosody to speech perception becomes more important when the segmental quality of speech or the listening conditions become less favorable.

Acknowledgments

I am most grateful for critical comments on earlier drafts of this chapter by René Collier from IPO, Eindhoven, and Robert D. Ladd from the Department of Linguistics, University of Edinburgh. The following people helped me in several practical ways to put this chapter into shape: Leo Vogten from IPO, Eindhoven; Eleonora Blaauw, Guus de Krom, and Hugo Quené from OTS, Utrecht.

References:

Abel, S.M. (1972a). Duration discrimination of noise and tone bursts. Journal of the Acoustical Society of America, 51, 1219-1223.

Abel, S.M. (1972b). Discrimination of temporal gaps. Journal of the Acoustical Society of America, 52, 519-524.

Adriaens, L.M.H. (1991). Ein Modell deutscher Intonation. Unpublished Doctor's Thesis. Eindhoven: Technical University of Eindhoven.

Allen, J., M.S. Hunnicutt, D. Klatt (1987). From Text to Speech: the MITalk System. Cambridge, MA: M.I.T Press.

Atal, B.S. and S.L Hanauer (1971). Speech analysis and synthesis by linear prediction of the speech wave. Journal of the Acoustical Society of America, 50, 637-655.

Baart, J.L.G. (1987). Focus, Syntax, and Accent Placement. Unpublished Doctor's Thesis. Leyden: Leyden University.

Barnwell, T.P. (1971). An Algorithm for Segment Durations in a Reading Machine Context. Technological Report No. 279, Research Laboratory of Electronics. Cambridge MA: Massachusetts Institute of Technology.

Bertinetto, P. (1981). Strutture Prosodiche dell'Italiano. Firenze: Academia della Crusca.

Blaauw, E. (1994) The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. To appear in Speech Communication 14(4).

Bolinger, D.L. (1958). A Theory of Pitch Accent in English. Word, 14, 109-149.

Borden, G. and K.S. Harris (1983). Speech Science Primer: Physiology, Acoustics and Perception of Speech. Baltimore, London: Wilkins and Wilkins.

Brokx, J.P.L. and S.G.Nooteboom (1982). Intonation and the perceptual separation of simultaneous voices. Journal of Phonetics, 10, 23-36.

Brown, G., K.L. Currie and J. Kenworthy (1980). Questions of Intonation. London: Croom Helm.

Bruce, G. (1977). Swedish Word Accents in Sentence Perspective. Travaux de l'Institut de Linguistique de Lund. Lund: CWK Gleerup.

Burghardt, H. (1973a). Die subjektive Dauer schmalbandiger Schalle bei verschiedenen Frequenzlagen. Acustica, 28, 278-284.

Burghardt, H. (1973b). Über die subjektive Dauer von Schallimpulsen und Schallpausen. Acustica, 28, 284-290

Campbell, N. (1990). Evidence for a syllable-based model of speech timing. Proceedings of the First International Congress of Spoken Language Processing (pp. 9-12). Kobe: Acoustic Society of Japan.

Campbell, N. (1992). Segmental elasticity and timing in Japanese speech. In Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (eds.), Speech Perception, Production and linguistic Structure (pp. 403-418). Tokyo, Osaka, Kyoto: Ohmsha; Amsterdam, Washington, Oxford: IOS Press.

Campbell, W.N. and S.D. Isard (1991). Segment durations in a syllable Frame. Journal of Phonetics, 19, 37-47.

Carlson, R. and B. Granström (1986). Linguistic processing in the KTH multi-lingual Text-to-Speech System. Proceedings ICASSP 1986, 2403-2406. Tokyo.

Charpentier, F. and E. Moulines (1989). The pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Proceedings of the european Conference on Speech Communication and Technology, Vol. II, 13-19. Paris.

Cherry, E.C. (1953). Some experiments on the recognition of speech, with one and with two ears. Journal of the Acoustical Society of America, 25, 975-979.

Chomsky, N. and M. Halle (1968). The Sound Pattern of English. New York: Harper and Row.

Collier, R. (1975). Perceptual and linguistic tolerance in intonation. International Review of Applied Linguistics, 13, 293-308.

Collier, R. and J.'t Hart (1972). Perceptual experiments on Dutch intonation. In A. Rigault and R. Charbonneau (eds.), Proceedings of the seventh international Congress of phonetic Sciences, pp. 880-884. The Hague, Paris: Mouton.

Cooper, W.E. and J. Paccia-Cooper (1980). Syntax and Speech. Cambridge MA, London UK: Harvard University Press.

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

Crystal, D. (1972). The Intonation System of English. In D.L. Bolinger (ed.), Intonation (pp. 110-136). Middlesex UK: Penguin, Harmondsworth.

Crystal, T.H. and A.S. House (1982). Segmental durations in connected-speech signals: preliminary results. Journal of the Acoustical Society of America, 72, 705-716.

Crystal, T.H. and A.S. House (1988a). Segmental durations in connected-speech signals: current results. Journal of the Acoustical Society of America, 83, 1553-1573.

Crystal, T.H. and A.S. House (1988b). Segmental durations in connected-speech signals: syllabic stress. Journal of the Acoustical Society of America, 83, 1574-1585.

Crystal, T.H. and A.S. House (1988c). The duration of American English vowels: an overview. Journal of Phonetics, 16, 263-284.

Crystal, T.H. and A.S. House (1988d). The duration of American English stop consonants: an overview. Journal of Phonetics, 16, 285-294.

Crystal, T.H. and A.S. House (1990). Articulation rate and the duration of syllables and stress groups in connected speech. Journal of the Acoustical Society of America, 88, 101-112.

Darwin, C.J. (1975). On the dynamic use of prosody in speech perception. In A. Cohen and S.G. Nooteboom (eds.), Structure and Process in Speech Perception (pp. 178-193). Berlin: Springer Verlag.

Dirksen, A. and H. Quené (1993). Prosodic analysis: the next generation. In V.J. van Heuven and L.C.W. Pols (eds.), Analysis and Synthesis of Speech (pp. 131-146). Berlin, New York: Mouton de Gruyter.

Eefting, W.Z.F. (1991). Timing in Talking. Tempo Variation in Production and its Role in Perception. Unpublished Doctor's Thesis. Utrecht: Utrecht University.

Fant, G. and A. Kruckenberg (1988a). Contributions to temporal analysis of read Swedish. Department of Linguistics, Working papers No. 34 (pp. 37-41). Lund: Lund University.

Fant, G. and A. Kruckenberg (1988b). Some durational correlates of swedish prosody. Proceedings of the seventh FASE Symposium, Vol. 2 (pp. 495-503). Edinburgh.

Fant, G. and A. Kruckenberg (1989). Preliminaries to the study of Swedish prose reading style. Speech Transmission Laboratory, Quarterly Progress Report No 2/1989 (pp. 1-83). Stockholm: Royal Institute of Technology.

Fant, G., L. Nord and A. Kruckenberg (1986). Individual variations in text reading. A Data Bank Pilot Study. Speech Transmission Laboratory, Quarterly Progress Report No. 4/1986 (pp. 1-7). Stockholm: Royal Institute of Technology.

Fant, G., L. Nord and A. Kruckenberg (1987). Segmental and prosodic variabilities in connected speech. An applied Data Bank Study. Proceedings of Eleventh International Congress of Phonetic Sciences, Vol. 6 (pp. 102-105. Tallinn: Estonian Academy of Sciences.

Flanagan, J.L. and M.G. Saslow (1958). Pitch discrimination for synthetic vowels. Journal of the Acoustic Society of America, 30, 435-442.

Fujisaki, H., K. Nakamura, T. Imoto (1973). Auditory Perception of Duration of Speech and Nonspeech Stimuli. Annual Report of Engineering Research Institute (p. 32). Tokyo: Faculty of Engineering, University of Tokyo

Fujisaki, H. and H.Sudo (1971). Synthesis by rule of prosodic features of connected Japanese. Proceedings of the Seventh International Congress on Acoustics, Vol. 3, 133-136. Budapest: Akadémiai Kiadó.

Geluykens, R. (1987). Intonation and speech act type. An experimental approach to rising intonation in queclaratives. Journal of Pragmatics, 11, 483-494.

Goldman-Eisler, F. (1968). Psycholinguistics: Experiments in spontaneous Speech. New York: Academic Press.

Goldstein, J.L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. Journal of the Acoustical Society of America, 54, 1496-1516.

Grosjean, F., L. Grosjean and H. Lane (1979). The patterns of silence: performance structures in sentence production. Cognitive Psychology, 11, 58-81.

Gussenhoven, C. (1983). Focus, mode and nucleus. Journal of Linguistics, 19, 377-417.

Gussenhoven, C. (1984). On the Grammar and Semantics of Sentence Accents. 1984: Foris Publications.

Hadding-Koch, K. and M. Studdert-Kennedy (1964). An experimental study of some intonation contours. Phonetica, 11, 175-185.

Hamon, C. (1988). Procédé en Dispositif de Synthèse de la Parole par Addition-Recouvrement de Formes d'Ondes. Patent no. 8811517.

Harris, M.S. and N. Umeda (1974). Effect of speaking mode on temporal factors in speech: vowel duration. Journal of the Acoustical Society of America, 56, 1016-1018.

Hart, J.'t (1976). Psychoacoustic backgrounds of pitch contour stylization. IPO Annual Progress Report, 11 (pp. 11-19). Eindhoven: Institute for Perception Research.

Hart, J. 't (1981). differential sensitivity to pitch distance, particularly in speech. Journal of the Acoustical Society of America, 69, 811-821.

Hart, J.'t and R. Collier (1975). Integrating different levels of intonation analysis. Journal of Phonetics, 1, 309-327.

Hart, 't J., R. Collier and A. Cohen (1990). A Perceptual Study of Intonation. An Experimental-phonetic Approach to Speech Melody. Cambridge, UK: Cambridge University Press.

Hermes, D.J. and Van Gestel, J.C. (1991). The frequency scale of speech intonation. Journal of the Acoustical of America, 90, 97-102.

House, A.S. (1961). On vowel duration in English. Journal of the Acoustical Society of America, 33, 1174-1178.

Huggins, A.W.F. (1968). The perception of timing in natural speech: compensation within the syllable. Language and Speech, 11, 1-11.

Huggins, A.W.F. (1972). Just noticeable differences for segment durations in natural speech. Journal of the Acoustical Society of America, 51, 1270-1278.

Huggins, A.W.F. (1975). Temporally segmented speech. Perception and Psychophysics, 18, 149-157.

Klatt, D.H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual Evidence. Journal of the Acoustical Society of America, 59, 1208-1221.

Klatt, D.H. and W.E. Cooper (1975). Perception of segment duration in sentence contexts. In A. Cohen and S.G. Nooteboom (eds.), Structure and Process in Speech Perception (pp. 69-89). Berlin: Springer Verlag

Kozhevnikov, V.A. and L.A. Chistovich (1965). Speech: Articulation and Perception (trans. U.S. Department of Commerce, Clearing House for Federal Scientific and Technical Information). Washington D.C.: Joint Publications Research Service.

Kulas, W. and H.W. Rühl (1985). Syntex - unrestricted conversion of text to speech for German. In R. de Mori and C.Y. Suen (eds.), New Systems and Architectures for automatic Speech Recognition and Synthesis (pp. 517-535).

Ladd, D.R. (1980). The Structure of intonational Meaning: Evidence from English. Bloomington IN: Indiana University Press.

Ladd, D.R. (1988). Declination “reset”and the hierarchical organization of utterances. Journal of the Acoustical Society of America, 84, 530-544.

Ladd, D.R. and K.E.A.Silverman (1984). Vowel intrinsic pitch in connected speech. Phonetica, 41, 31-40.

Lehiste, I. (1970), Suprasegmentals. Cambridge MA, London UK: The M.I.T. Press.

Liberman, M.Y. (1979). The intonational System of English. New York: Garland.

Liberman, M. and Pierrehumbert, J. (1984), Intonational invariance under changes of pitch range and length. In: M.Aronoff and R. Oehrle (eds.), Language and Sound Structure. Cambridge, MIT Press, pp. 157-233.

Lindblom, B.E.F. (1968). Temporal organization of syllable production. Speech Transmission Laboratory, Quarterly Progress Report No 2-3/1968 (pp. 1-5). Stockholm: Royal Institute of Technology.

Lindblom, B.E.F. (1989). Phonetic invariance and the adaptive nature of speech. In B.A.G. Elsendoorn and H. Bouma (eds.), Working Models of human Perception (pp. 139-173). London, San Diego, New York, Berkeley, Boston, Sydney, Tokyo, Toronto: Academic Press.

Lindblom, B. and K. Rapp (1973). Some temporal regularities of spoken Swedish. PILUS 21 (Papers from the Institute of Linguistics). Stockholm: University of Stockholm.

Lisker, L. (1957). Closure duration and the intervocalic voiced voiceless distinction in English. Language, 33, 42-49.

Maeda, S. (1976). A Characterization of American English Intonation. Unpublished PhD thesis. Cambridge, MA: MIT.

Murray, I.R. and J.L. Arnott (1993). Toward the simulation of emotion in synthetic speech: A Review of the Literature on human vocal Emotion. Journal of Acoustical Society of America, 93, 1097-1108.

Nakatani, L.H. and J.A.Schaffer (1978). Hearing “words” without words. Journal of the Acoustical Society of America, 63, 234-245.

Nespor, M. and M. Vogel (1986). Prosodic Phonology. Dordrecht: Foris Publications.

Nooteboom, S.G. (1972). Production and Perception of Vowel Duration. A Study of durational Properties of Vowels in Dutch. Unpublished Doctor’s Thesis. Utrecht: University of Utrecht.

Nooteboom, S.G. (1973). The perceptual reality of some prosodic durations. Journal of Phonetics, 1, 25-45.

Nooteboom, S.G. (1979). “Time” in the production and perception of speech. Arbeitsberichte nr 12: Report of an interdisciplinary Colloquium held in the Phonetics Department of Kiel University, February 22-24, 1979 (pp. 113-151). Kiel: Institut für Phonetik, Universität Kiel.

Nooteboom, S.G. (1985). A functional view of prosodic timing. In J.A. Michon and J.L. Jackson (eds.), Time, Mind, and Behavior (pp. 242-251). Berlin, Heidelberg, New York, Tokyo: Springer-Verlag.

Nooteboom, S.G., J.P.L.Brokx and J.J.De Rooij (1978). Contributions of prosody to speech perception. In W.J.M.Levelt and G.B. Flores d’Arcais (eds.), Studies in the Perception of Language (pp. 75-107). New York: Wiley.

Nooteboom, S.G. and G.J.N. Doodeman (1980). Production and perception of vowel length in spoken sentences. Journal of the Acoustical Society of America, 67, 276-287.

Nooteboom, S.G. and J.G. Kruyt (1987). Accents, focus distribution, and the perceived distribution of given and new Information: an experiment. Journal of the Acoustical Society of America, 82, 1512-1524.

Odé, C. (1989). Russian Intonation: a Perceptual Description. Amsterdam, Atlanta: Rodopi.

Ohde, R.N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. Journal of the Acoustical of America, 75, 224-230.

Os, E.A. den (1988). Rhythm and Tempo of Dutch and Italian. Unpublished Doctor's Thesis. Utrecht: Utrecht University.

O'Shaughnessy, D. (1976). Modelling Fundamental Frequency, and its Relationship to Syntax, Semantics, and Phonetics. Unpublished PhD thesis. Cambridge, MA: MIT.

O'Shaughnessy, D. (1979). Linguistic features in fundamental frequency patterns. Journal of Phonetics, 7, 119-145.

O'Shaughnessy, D. (1987). Speech Communication, Human and Machine. Reading MA, Menlo Park CA, New York, Don Mills, Wokingham UK, Amsterdam, Bonn, Sydney, Singapore, Tokyo, Madrid, Bogotá, Santiago, San Juan: Addison-Wesley Publishing Company.

Peterson, G.E. and H.L. Barney (1952). Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 24, 175-184.

Pierrehumbert, J. (1980). The Phonology and Phonetics of English Intonation. Unpublished PhD thesis. Cambridge, MA: MIT.

Pijper, J.R. de (1983). Modelling British-English Intonation. Dordrecht, Cinnaminson: Foris Publications.

Quené, H. (1992). Durational cues for word segmentation in Dutch. Journal of Phonetics, 20, 331-350.

Quené, H. and R. Kager (1993). Prosodic sentence analysis without parsing. In V.J. van Heuven and L.C.W. Pols (eds.), Analysis and Synthesis of Speech (pp. 115-130). Berlin, New York: Mouton de Gruyter.

Rietveld, A.C.M. and C. Gussenhoven (1985). On the relation between pitch excursion size and prominence. Journal of Phonetics, 13, 299-308.

Ritsma, R.J. (1967). Frequencies dominant in the perception of the pitch of complex sounds. Journal of the Acoustical Society of America, 42, 191-199.

Rooij, J.J. de (1978). Speech Punctuation. An Acoustic and Perceptual Study of some Aspects of Speech Prosody in Dutch. Unpublished Doctor's Thesis. Utrecht: University of Utrecht.

Ruhm, H.B., E.O. Mencke, B. Milburn, W.A. Cooper and D.E. Rose (1966). Differential sensitivity to duration of acoustic signals. Journal of Speech and Hearing Research, 9, 371-384.

Santen, J.P.H. van (1992). Contextual effects on vowel duration. Speech Communication, 11, 513-546.

Santen, J.P.H. and J.P. Olive (1989). The Analysis of contextual Effects on segmental Duration. Computer, Speech and Language, 4, 359-390.

Schouten, J.F. (1940). The Residue and the Mechanism of Hearing. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen 43, 991-999.

Selkirk, E. (1984). Phonology and Syntax: the Relation between Sound and Structure. Cambridge MA: M.I.T. Press.

Silverman, K.E.A. (1986) F_0 segmental Cues depend on
intonation: the case of the rise after voiced stops. Phonetica, 43, 76-91.

Slis, I.H. and A. Cohen (1969). On the complex regulating the voiced-voiceless distinction, I and II. Language and Speech, 12, 80-102 and 137-156.

Slootweg, A. (1988). Metrical prominence and syllable duration. In P. Coopmans and A. Hulk (eds.), Linguistics in the Netherlands 1988 (pp. 139-148). Dordrecht: Foris Publications.

Steele, S.A. (1986). Interaction of vowel F_0 and prosody. Phonetica, 43, 92-105.

Terken, J.M.B. and S.G. Nooteboom (1987). Opposite effects of accentuation and deaccentuation on verification latencies for given and new Information. Language and Cognitive Processes, 2, 145-163.

Thorsen, N. (1980). A study of the perception of sentence intonation - Evidence from Danish. Journal of the Acoustical Society of America, 67, 1014-1030.

Thorsen, N. Gronnum (1985). Intonation and text in standard Danish. Journal of the Acoustical Society of America, 80, 1205-1216.

Willems, N.J., J.'t Hart and R. Collier (1988). English intonation from a Dutch point of view. Journal of the Acoustical Society of America, 84, 1250-1261.

Captions

Fig. 1: Measured course of pitch in a Dutch sentence, with only two voluntary pitch movements, an accentuating rise on the syllable “KLEIN” and an accentuating fall on the syllable “DENT”. All other pitch movements are involuntary side-effects of other speech processes. Note also that the continuity of pitch is interrupted during all voiceless consonants.

Fig. 2: Measured course of pitch (dotted line) in a British-English utterance together with a so-called “close-copy” stylization (interrupted line), containing the smallest possible number of straight-line segments with which perceptual equality between original and close-copy can be achieved.

Fig. 3: Measured course of pitch (dotted line), close-copy stylization (interrupted line), a grid of three declination lines (solid lines), and standardized stylization (bold line) in a British-English utterance.

Fig. 4: Oscillographic representation of the utterance “The queen said, the knight is a monster”.

Fig. 5: Schematized temporal patterns of reiterant versions of Dutch spoken words with stress on the first syllable, and varying from one to four syllables. Top: with repetitions of the syllable [ma:m], bottom: with repetitions of the syllable [m✌m].

Fig. 6: Calculated (solid line), spoken (crosses) and adjusted (circles) durations of stressed [a:] (top curve) and [✌] (bottom curve) as a function of the number of syllables in the word which remain to be produced at the beginning of the syllable concerned.

Fig. 7: Three oscillograms. Top: an originally spoken Dutch utterance. Bottom: a reiterant version of the same utterance, each syllable being spoken as [ma:]. Bottom: Same as top, but with syllable durations made identical to those in the reiterant version.

Fig. 8. Two oscillograms. Top: “The queen said, the knight is a monster”. Bottom: “The queen, said the knight, is a monster”.