1

Pitch in Humans and Machines

1.1 Introduction

In this first chapter, some phonetic information is given which will be required for an active engagement in research in the area of tone and intonation. For further information on the articulatory, acoustic, and technological facts, handbooks like Laver (1990), Ladefoged (1996), Johnson (1997), Rietveld and van Heuven (1997), Reetz (1999) should be consulted.

1.2 Frequency of vocal fold vibration, fundamental frequency (F_0), and pitch

Pitch is the auditory sensation of tonal height. We have this sensation when listening to the difference between [s] and [ʃ], for instance, but in speech, it is most precise when it reflects the periodicity in the acoustic signal. Periodicity amounts to repetitions of the same pattern of vibration, each such repetition being a *period* and corresponds to a closing-and-opening action of the vibrating vocal folds. The actual shape of the speech signal during a period determines the sound quality (the vowel quality, say) that we perceive. In panel (a) of figure 1.1, 25 milliseconds (ms) from the speech waveform produced by a woman are shown. During that time, just over six periods were produced, representing as many vibratory cycles of the vocal folds. These are two muscles, situated halfway down the larynx, which run front to back from the inside of the thyroid (the shield cartilage sticking out in the front of the neck) to the two arytenoids, which are located above the cricoid. In a relaxed state, there tends to be a slit between the vocal folds, known as the glottis, through which we breathe. The slit can be widened to a triangular opening by rolling the arytenoids away from each other, allowing increased air intake or escape. They can also be brought more closely together, to a point where their edges will be pushed up by the air pressure below them during the exhalation phase, prising the glottis open. The subsequent drop

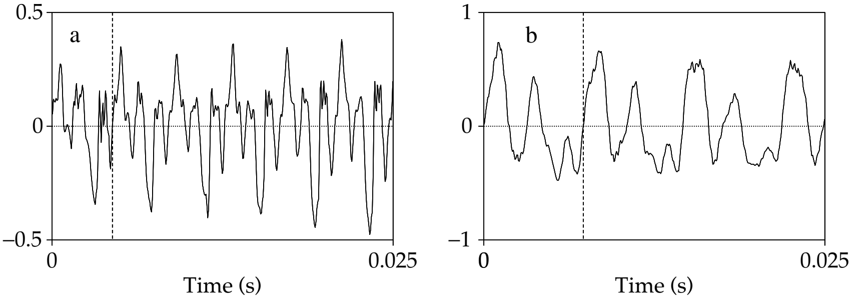

Fig. 1.1 Sections of 25 ms from the speech waveform during vocal fold vibration produced by a woman (panel a) and produced by a man (panel b). The vertical dotted lines mark of the end of the first period, which is 4.42 in panel (a) and 7.28 ms in panel (b).

in air pressure between the vocal folds, which is due to the rapid flow of air through the glottis (the Bernoulli effect), will cause them to be sucked together again, after which the subglottal pressure will again push them up, and apart, and so on. Vocal fold vibration is the alternation between these opening and closing events.

The opening action of the vocal folds is normally slower than the snappier closing action. Each closing action has an effect which is comparable to that of the flick of a finger against the throat: a brief shock wave is set up, which hits the walls of pharynx and mouth. If this happens more than forty times a second, we stop hearing the flicks as separate events, and instead perceive a continuous event: pitch. The faster these waves follow each other, i.e. the higher the frequency of vibration of the vocal folds, the higher will be the number of periods per second, commonly known as the *fundamental frequency*, or F_0 (‘F-zero’), of the acoustic signal, and the higher the resulting pitch. The beginning of the graph in panel (a) coincides with a point in time during which the air pressure is neither raised nor lowered as compared to the surrounding air pressure, a zero-crossing. The first period ends at the dotted vertical line, which has been drawn through a zero-crossing. The duration of this period is 4.42 ms. As can be seen, the waveform during this period contains three positive peaks, of which the second and third clearly reveal further smaller peaks within them. These higher frequencies ride on the crest of the F_0, and in this case cause it to be heard as a mid front vowel. Each of the six closing actions of the vocal folds that are responsible for the six periods in panel (a) of figure 1.1 is thus comparable to what happens when you flick a finger against the skin of the neck, near the larynx. When this is done while holding the larynx closed and holding the mouth in the position for some vowel, [ɛ] in this case, a popping noise is produced, which rapidly fades away.[1]

F_0 is usually expressed in Hz (or hertz), the number of periods per second. The F_0 corresponding to the first period in panel (a) is 1,000 (ms) divided by

4.42 ms, the period, or 226 Hz. In panel (b) of figure 1.1, 25 ms from a speech waveform produced by a man are shown. Its shape during the period corresponds to that of high back vowel, [u]. As will be clear, only about 3.5 periods fit into this time span. The first period is 7.26 ms long, and the F_0 is therefore 138 Hz. Rates of vibration in male speakers average around 125 Hz while those in female speakers, whose larynxes are much smaller in the front to back dimension than those of men, average around 225 Hz (cf. Holmberg, Hillman, and Perkell 1988).

1.3 Pitch tracks

Several techniques are available for recording the pitch of utterances (Hermes 1993; Reetz 1996: 83ff.). They can be based on the articulation, the acoustics, or the perception. By definition, the best source for obtaining a record is the listeners' perception, since pitch is a perceptual sensation. Unfortunately, listeners lack the appropriate conceptualizations and vocabulary to report their sensations, and are typically incapable of saying even whether a given pitch change represents a fall or a rise. Things become very much simpler if the voiced parts of the speech signal are divided up into sections of some 30 ms, and static pitch judgements are obtained for each of these, as recorded on a scale from low to high. The evidently laborious nature of this procedure is not its only drawback. It is still unreliable because of the inaccurate way in which people report their pitch sensations, a drawback which can only be overcome with the help of a careful experimental design allowing averaging over many trials. Not surprisingly, this method of reporting pitch variation is rarely, if ever, used.

An effective way of measuring the vocal fold vibration is by running a weak electric current between two electrodes placed on the skin on either side of the larynx, so that the vocal folds lie between them. Because the impedance for the electric signal emitted by the first electrode is increased when the vocal folds open, the opening actions can be read off the electric signal reaching the second electrode (Fourcin and Abberton 1971). While it generally records vocal fold vibration rates accurately, this method can be used only in laboratory conditions.

Most commonly, records are obtained from the speech signal. There are many ways in which the F_0 of a signal can be established automatically, which are commonly referred to as 'pitch trackers' (even though they measure the fundamental frequency). The evaluation of their merits is virtually a field of study in itself (cf. Reetz 1999, who refers to Hess 1983, Hermes 1993, and Reetz 1996). Because they are implemented as computer algorithms, pitch trackers need to work with digitized forms of the signal. Digitization of the continuous speech waveform is performed by determining a single value at regular intervals, as shown in panel (a) in figure 1.2. The number of times per second that a measurement of the waveform is recorded is the *sampling rate*. A musical CD-Rom contains 44.100 measurements per second, and thus has a sampling rate of 44.1 kHz, a DAT-recording stores the signal at a 48 kHz sampling rate, while many speech scientists make do with a sampling rate 16 kHz. Panel (b) shows

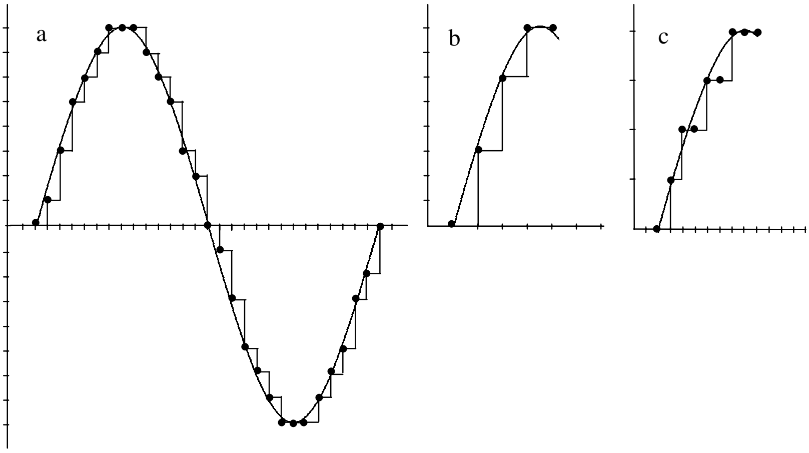

Fig. 1.2 A continuous speech waveform and a digitized waveform (panel a), and digitized waveforms with a lower sampling rate (panel b), and a lower quantization accuracy (panel c).

part of the same waveform sampled with half the sampling rate of that in panel (a), as is evident from the calibration marks on the time axis and the corresponding durations of the 'steps'. As a result, the first upward excursion which in panel (a) is represented by seven values (including the first sample with zero excursion), is represented by only four values in panel (b). A second parameter determining the reproduction quality of the digitized signal is the accuracy with which the value of each sample is stored, or *quantized*. The lower the accuracy, expressed in bits,[2] the larger the rounding errors, and the greater the jumps from one sample to the next will be. This is illustrated in panel (c). Due to the larger rounding errors, instead of the seven values of panel (a), there are effectively only five values, even though the signal is sampled at the same rate. Larger jumps lead to more *quantization noise*. A 16-bit accuracy is a good choice for music, while 8-bit accuracy is not uncommon in speech research.

It stands to reason that in order to preserve a period in a digitized signal, at least two values for that period must be recorded. If the sampling rate were to equal the signal frequency, the information would be reduced to a single value. This is the reason why the sampling rate must be (preferably more than) twice the highest signal frequency we wish to preserve. Signal frequencies that lie between half the sampling rate and the sampling rate give rise to 'false' periods. This 'aliasing' effect is comparable to the visual effect of spoked wheels turning backwards in cinefilms. This is why all signal frequencies above half the sampling rate are filtered out before digitization. A sampling rate of 16 kHz is therefore way below what would be required for hifi equipment, since younger people may perceive frequencies of up to 20 kHz. However, it is usually good enough for speech, where no significant phonetic information is found above 8 kHz.

Once the signal is represented as a digital file, automatic F_0 detection can begin. Broadly, a division can be made between algorithms that operate directly on the digitized speech signal and those that take a spectral analysis as input. Algorithms that work directly on the signal employ some form of *autocorrelation*. The series of values within an analysis window, whose duration might be chosen to fall well above the period to be detected, is established, to represent the window's starting position. As the window is moved through the speech waveform, sample by sample, the string of values encountered at each step is correlated with those of the starting position. Clearly, the sections of the waveform within the window encountered at each step are not likely to resemble that in the starting position at all closely, except when the window begins at an equivalent point in a following period. At that point, the sections will be very similar, resulting in a high positive correlation between the strings of values. Alternatively, the duration of the window is gradually increased from a starting duration well below the estimated period duration, and the values in each of the incremented windows is correlated with the values in the next section having the same duration. Again, when the duration of the window equals the duration of the period, the correlation will be relatively high, reflecting the fact that the values in the two periods are now very similar. In practice, the choice of pitch tracker is determined by circumstance and convenience: their evaluation is typically beyond the competence of a phonologist.

1.4 Interpreting pitch tracks

The results of a pitch tracker can be plotted against time to provide a visual representation of the F_0 variation in the utterance. The x-axis gives time, usually in milliseconds (ms), the y-axis F_0, commonly reported in hertz (Hz). It is sometimes felt that a linear representation of the number of periods per second gives a biased view of the auditory impression of pitch changes. For instance, a difference between 600 Hz and 650 Hz seems smaller than one between 100 and 150. Two measures that have been claimed to give a better auditory representation are the semitone scale (ST) and the Equivalent Rectangular Bandwidth (ERB) scale. According to Hermes and van Gestel (1991), of the three scales this scale represents auditory impressions in intonation most closely. In practice, the choice here is hardly ever a problem, since within the usual F_0 range in speech, ERB and Hz values are very similar (Rietveld and van Heuven 1997: 210). In this book, Hz scales are used.

1.4.1 Tracking errors

Pitch trackers make mistakes. A pitch tracker may mistake voiceless friction for voicing, and irregular measurement points may therefore show up during fricatives and releases of affricated plosives. Conversely, it may interpret irregularly

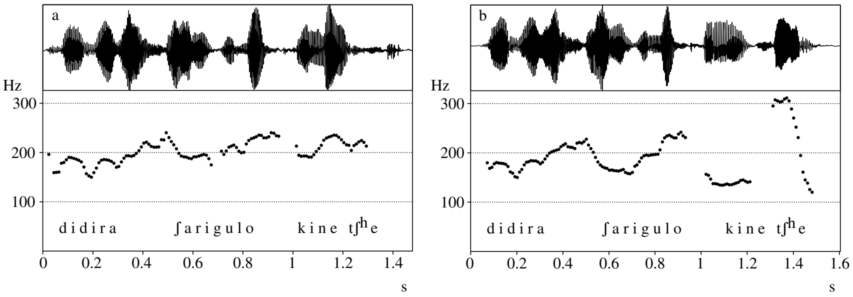

Fig. 1.3 Incorrectly detected voicing during the fricative part of an affricate and incorrectly detected voicelessness during creaky [e] in the last syllable of Bengali ‘My sisters (also) BOUGHT the saris’ (panel a) and a correctly analysed contour for an interrogative pronunciation of the same sentence, ‘Did my sisters buy the saris?’ (panel b).

voiced signals as voicelessness. Of course, the researcher can always inspect the waveform and measure the period(s) he or she is interested in so as to calculate the corresponding F_0 by hand. Both problems appear in the pitch track in figure 1.3. Incorrectly detected voicing during the aspirated affricate tʃʰ occurs at 1.3 s, while incorrectly detected voicelessness during creaky [e] occurs at 1.4 s. The F_0 of this vowel is actually around 120 Hz, but no record was made. Finally, there may be periodic background sounds that are picked up by the pitch tracker, causing it to report spurious F_0 values. For comparison, consider the contour in panel (b), which represents a different intonation pattern, and where the tʃʰe has been correctly analysed as voicelessness followed by a steep fall on [e].

Even if the detection of voicing and voicelessness is correct, the pitch tracker may fail to analyse the voiced signal correctly. When the voice becomes creaky, as it often does at lower pitches, the algorithm may be confused by peaks in the signal that do not correspond to the vibratory action of the vocal folds. If these appear halfway through the period, they may be interpreted as peaks created by the opening actions of the vocal folds, leading to fundamental frequency measurements that are twice that of the ‘real’ fundamental frequency (*doubling errors*). Similarly, the algorithm may miss every second periodicity peak, ‘believing’ these peaks determine the sound quality rather than the periodicity (*halving errors*). Such ‘octave jumps’ are usually easy to detect: the pitch track shows a sudden change to a value half or double that of the immediately preceding value, while there is no auditory impression that corresponds to this jump. The pitch track in figure 1.4 shows a halving error in the last part of the vowel [ɪ]. A corrected version with different analysis settings, obtained by trial and error, appears in panel (c) of figure 2.2. Among other things, pitch trackers allow the user to choose the range of possible frequencies to be detected. If the actual F_0 of the speech file falls outside the upper or lower limits of the analysis band, mistakes are inevitable. Safe ranges are usually 75–400 Hz for male speakers

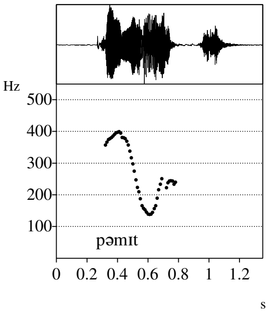

Fig. 1.4 Halving error in last part of the stressed vowel of *permit?* (verb). The pitch continues upwards to 480 Hz, but the F_0 tracker shows a record around 240 Hz.

and 100–600 Hz for female speakers. Children may produce more than 600 Hz. (I have not corrected the graphic results of the pitch analyses reproduced in this book.) Errors other than octave jumps will also occur.

1.4.2 Consonantal effects on F_0

The articulation of segments will interfere with the production of vocal fold vibration. First, if all is well, pitch trackers will report no F_0 during glottal closures or voiceless consonants like [p,s,χ,m̥], there being no vocal fold vibration. Due to their oral constrictions, voiced obstruents, too, particularly plosives like [b,d,g], may impede the airflow needed to keep the vocal folds vibrating, slowing the vibration down or stopping it altogether. By contrast, sonorant consonants and vowels allow the air coming in from the lungs to escape via the nostrils or the opened lips to prevent air pressure from building up in the mouth, and the vocal fold vibration is therefore uninhibited during these sounds. Because our pitch perception in language normally factors such effects out, what is subjectively the same pitch contour may look rather different depending on the consonants in them. This is illustrated in panel (a) in figure 1.5, which gives F_0 tracks of the same intonational fall produced on the segmental structures [ata], [ada], and [ana].

The production of a voiceless consonant requires an active gesture of the arytenoids opening the glottis, but the cricothyroid muscle is also active (Löfqvist, Baer, McGarr, and Story 1989). This may cause the vocal folds to be somewhat tighter than during relaxed vibration. The effect is typically still present during the vibration for the following vowel, causing the fundamental frequency after voiceless consonants to be higher than after voiced consonants. A particularly clear example of the effect is shown in panel (b) of figure 1.5, where a consonantal 'pitch perturbation' occurs after the [ks] of *niks* in the Dutch utterance *Ik kan gewoon niks anders* 'I can't do anything else.' There is a raised F_0 after the [ks] of *niks*. Again, the auditory impression is that of a smoothly falling pitch

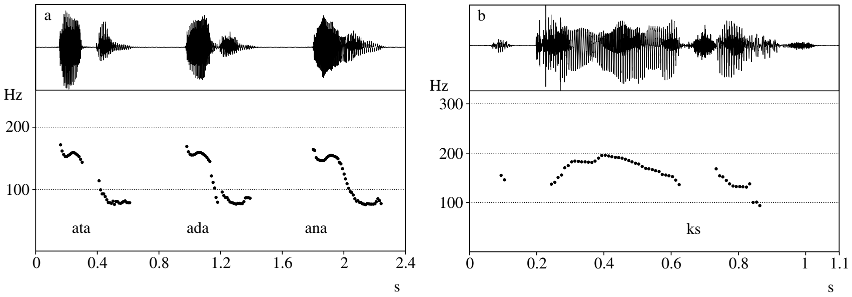

Fig. 1.5 Pitch falls in VCV-structures with a voiceless obstruent, a voiced obstruent, and a sonorant consonant for C, respectively (panel a), and an F_0 perturbation after [ks] in Dutch (panel b).

contour. Although this is generally hard to detect in F_0 tracks, a small effect must also exist *before* obstruents. Some lowering occurs for both voiced and voiceless consonants, but the lowering is greater before voiced consonants (Silverman 1984; Silverman 1990). The F_0 in the section of the vowel immediately preceding [t] and [d] was shown to influence the perception of the voicing of the consonant in German and English (Kohler 1990): higher F_0 in the preceding vowel increased the chance that [t] was perceived in a task in which listeners could choose between [t] and [d].

Further effects are due to phonation type (Laver 1980). Breathy voice induces low pitch, although the mechanism is not entirely clear (Hombert, Ohala, and Ewan 1979). During breathy voice, more air escapes per opening action than is needed to keep the vibration going, the excess air being used to create friction in the glottal aperture. The combination of friction and vibration must be easier to obtain at lower frequencies. Laryngealized voice may be conducive to high pitch, since the vocal folds need to be stiffened to produce it. However, there is also the converse fact that creaky voice, a form of laryngealization which is produced with slacker vocal folds, is easier to obtain at lower frequencies (Kingston 2003).[3]

1.4.3 Vocalic effects on F_0

The articulation of vowels may also affect the tenseness of the vocal folds, and so interfere with the rate of vocal fold vibration. The most plausible explanation is that high vowels like [i,u] are pronounced with the tongue high in the mouth, causing the hyoid, the horseshoe-shaped bone to which the tongue root is attached, to pull up the forward part of the larynx, the thyroid, to which the vocal folds are attached. As a result, higher vowels will on average be pronounced with higher vibration rates than lower vowels, like [a]. The component in the F_0 which is due to the correlation between vowel height and F_0 is known as *intrinsic pitch*, but, as Reetz (1999) suggests, 'intrinsic F_0' would be a better term. The difference is larger in stressed than in unstressed syllables (Silverman 1990).

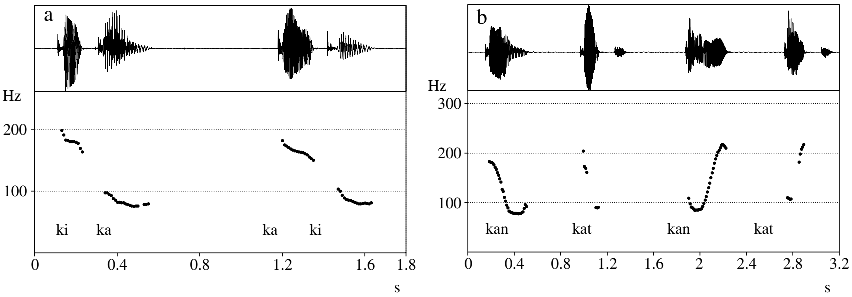

Fig. 1.6 The effect of intrinsic F_0 in acoustically different, but perceptually identical F_0 contours on [kika] and [kaki] (panel a), and fade-out reversals in [ken] compared with the same contours cut short by [t] in [kat] (panel b).

Panel (a) in figure 1.6 shows two subjectively identical pitch contours on [kika] and [kaki], respectively. However, the first syllable in the second word has lower F_0. Indeed, we do not normally hear intrinsic pitch as pitch. When [a] has the same fundamental frequency as [i], we hear it as higher than [i], and we thus factor out the effect in perception (Silverman 1985). In practice, intrinsic pitch need not cause any problems when relating the pitch track to our auditory impression of the pitch contour. However, when setting up experiments involving pitch, vowel height may have to be controlled for.

1.4.4 End-of utterance effects

Utterances ending in sonorant consonants (e.g. [m,r,l,w]) or vowels may end with a reversal of the F_0 in the last part of the utterance, where the signal fades out, a phase which may be detectable to the pitch tracker, but is ignored in human perception. There appears to be no discussion in the literature, but it would seem reasonable to assume that it is due to a relaxation of the muscles controlling the frequency of vibration of the vocal folds. Panel (b) in figure 1.6 shows a fall and a rise on the syllable [kɑn], with what might be called a fade-out reversal on each, i.e. a weak rise after the fall and a weak fall after the rise. When the same contours are spoken on the syllable [kɑt], the pitch movement occurs inside a high-intensity part of the signal. Here, the segment [t] cuts short movements, a phenomenon known as *truncation* or *curtailment*. (The pitch tracker had some problems in measuring the central portion of the vowels, as is evident from the gaps.)

Some languages truncate more drastically than others (Grabe 1998b). In German, final falls on rhymes consisting of a short vowel and a voiceless obstruent (e.g. *Schiff* 'ship') are quite drastically truncated, often leaving just a level portion. By comparison, the syllable with a long vowel, *schief* 'slanted'), or a disyllable like *Schiefer* (proper name), have contours that fall. Importantly, to the native ear, these intonation patterns are identical. In equivalent conditions in British English

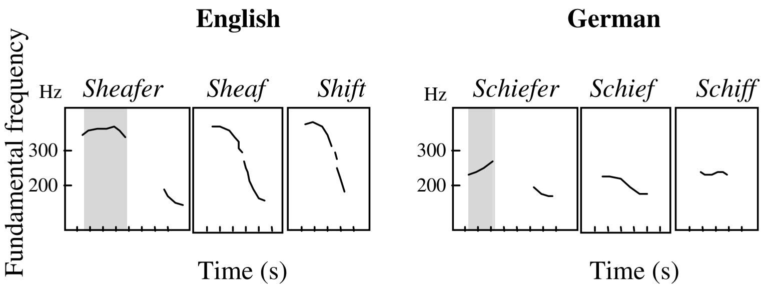

Fig. 1.7 Representative F_0 traces of falls in British English and German on a pre-final syllable, on a final syllable with a long vowel, and a final syllable with a short vowel followed by an obstruent. The shaded area indicates the stress vowel of a disyllabic word. German truncates falls on final rhymes like *-iff*, but English does not (cf. *shift*). From Grabe (1997: 163).

(e.g. *shift*), the contour is hardly different from what is found in segmentally more favourable conditions, like *Sheafer, Sheaf*, as shown in figure 1.7, from (Grabe 1998b). It is thus not only gaps that may 'distort' the visual picture but also early curtailments of the vocal fold vibration, something that may occur before final voiceless obstruents.

1.5 Experimentation

Our understanding of the prosodic structure of languages has greatly benefited from production and perception experiments. Production experiments may use either scripted or unscripted speech. Scripted speech is produced on the basis of written material which is read out or acted out in some fashion; while unscripted speech relies on spontaneous speech, elicited with the help of the Map Task or similar methods (Anderson *et al.* 1991). Such 'games' typically involve two speakers, who are drawn into conversations which stimulate the production of linguistic structures that the researcher is interested in, either because the speakers are invited to talk about items that have names exemplifying those structures or because communicative situations are created that are likely to lead to the production of particular intonation contours.

Perception experiments may use either natural stimuli, selected sections of naturally spoken speech, or stimuli in which the F_0 contour has been created artificially. Since 1995, an efficient technique has been available which performs such F_0 manipulation directly on the waveform (Moulines and Verhelst 1995). It replaces earlier techniques by which the specification of the F_0 component was altered after the signal had been analysed into the F_0 component and a number of spectral components, and a new signal was 'resynthesized' with the help of the components which included the altered values. The direct techniques reproduce shorter or longer versions of the periods detected in the signal by a procedure known as PSOLA (Pitch-Synchronous OverLapAdd). If the deviation