Measuring Speech. Fundamental frequency and pitch.

Daniel J. Hirst, Céline de Looze

In Rachael-Anne Knight and Jane Setter (eds) (in press)

The Cambridge Handbook of Phonetics.

Chapter 13. Fundamental Frequency and Pitch

Daniel Hirst and Céline De Looze

13.0 Abstract

In this chapter, we introduce the reader to the concepts of pitch and fundamental frequency from a functional, physiological and physical perspective. Several issues, including the modelling of intonation, pitch detection and measurement and acoustic scales, described below, are addressed to inform the reader about best practice for teaching and learning.

Pitch, corresponding to the subjective impression of whether individual speech sounds are perceived as relatively high or low, as on a musical scale, is an important characteristic of spoken language, contributing in some languages to the lexical identity of words (tone and accent) and in all languages to the perception of the intonation pattern of utterances. Pitch corresponds to the physiological parameter of the frequency of vibration of the vocal folds (aka vocal cords) which can be measured in cycles per second (cps), or the equivalent acoustic parameter of fundamental frequency (f0), measured in hertz (Hz).

Estimating and measuring fundamental frequency and modelling pitch is not an easy task. In this chapter, we first present some automatic models of pitch that have been developed both for speech synthesis and for the empirical study of intonation patterns. We then address issues related to the detection and measurement of fundamental frequency, including tracking/detection errors and explain how many of these errors can in fact be avoided by an appropriate choice of pitch ceiling and floor settings. We finally discuss the use of acoustic scales (e.g. linear, logarithmic, psychoacoustic) in the literature for the measurement of pitch. Based on evidence from recent findings in neuronatomy, neurophysiology, behavioural studies and speech production, we suggest that a new scale, the Octave-Median (OMe) scale, appears to be more natural for the study of speech prosody.

13.1 Introduction

Pitch is an important characteristic of spoken language, contributing in some languages to the lexical identity of words (via tone and accent) and in all languages to the perception of the intonation pattern of utterances. Pitch thus contributes in all languages to the interpretation of utterances, in ways which are not yet fully understood, via a number of different linguistic and

paralinguistic functions including the identification of speech acts (statement, question, command etc.), the recognition of different speaker states (attitudes and emotions etc.), the perception of prosodic structuring via prominence and phrasing, as well as many other discourse and dialogue related characteristics. See further Warren and Calhoun (this volume).

Most phoneticians make a distinction between *pitch* and *fundamental frequency*. The first corresponds to the subjective impression as to how voiced sounds, particularly sonorants and vowels, are perceived on a scale going from low to high as on a musical scale of notes, while the second corresponds to the physiological parameter of the frequency of vibration of the vocal folds (*aka* vocal cords) measured in cycles per second (cps), or the equivalent acoustic parameter of fundamental frequency (f_0), measured in hertz (Hz). The term pitch is, however, sometimes loosely applied to the acoustic or physiological measurement, as in the commonly used expressions *pitch detection* or *pitch range*.

13.2 Historical Overview

13.2.1 Pitch detection and analysis.

It was possible, even before the invention of speech recording, to make tracings of speech by means of a kymograph. This device, invented in the 1840s by the German physiologist Carl Ludwig, originally for monitoring blood pressure, was basically a revolving drum wrapped with a sheet of paper on which a stylus recorded changes in pressure as a function of time. Jones (1909) notes that this instrument had been used to make 'accurate records of intonation (…) by means of tracings of voice vibrations' (p iv).

With the invention of sound recording towards the end of the 19th century, it became possible, for the first time in history, to listen to utterances more than once.

Due to the laboriousness of using the kymograph and the difficulty of interpreting the output, Jones chose to use recordings from 'a Gramophone, Phonograph or other similar instrument' [p v], noting on a musical stave the pitch (or pitches) by picking up the needle immediately after hearing each syllable and identifying the corresponding musical note(s) of the syllable by comparison with a tuning fork.

Figure 13.1 gives an example of the type of transcriptions that Jones managed to produce using this technique:

In Rachael-Anne Knight and Jane Setter (eds) (in press)

The Cambridge Handbook of Phonetics.

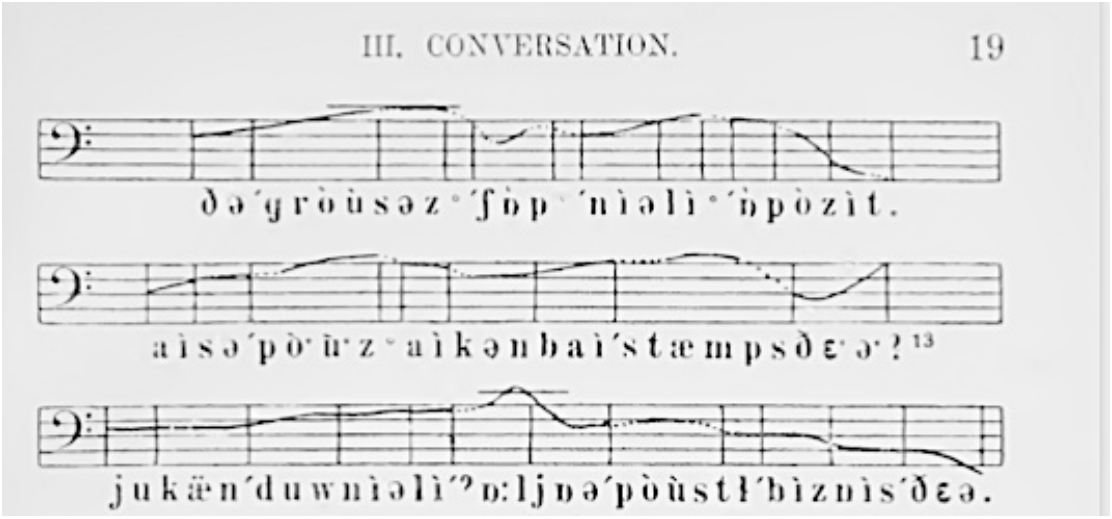


Figure 13.1. Sample transcriptions from Jones 1909. The sentences are: "The grocer's shop nearly opposite." "I suppose I can buy stamps there?" "You can do nearly all your postal business there".

In more recent research, speech recordings are used to produce an acoustic image by means of a computer. Figure 13.2 shows a portion of the waveform of a vowel /a:/ with one period highlighted, displayed using the Praat software (Boersma & Weenink, 1992). Here, the beginning and end points of the period have been chosen at a zero-crossing, although other choices are possible such as the maximum or minimum of the period. The representation of the speech waveform here is essentially the same as that obtained mechanically by a kymograph.

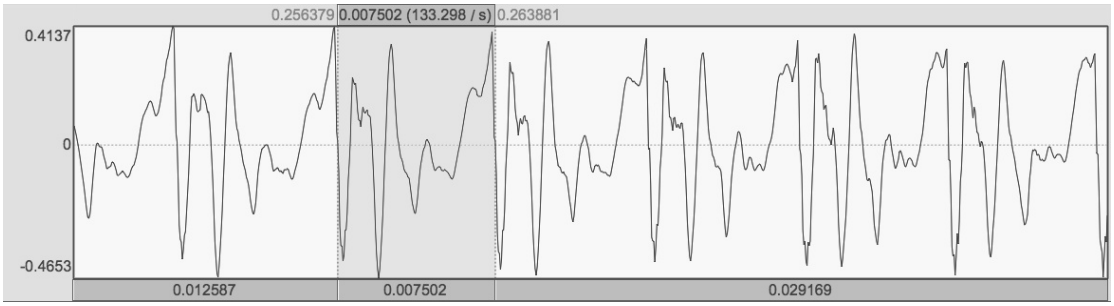
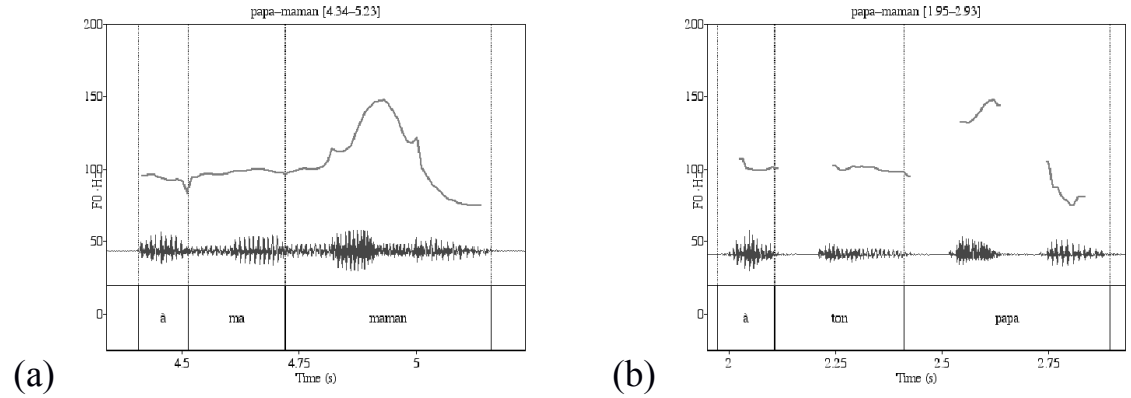


Figure 13.2 The waveform of a portion of a vowel /a:/ with one period highlighted.

From the waveform, the duration of the period can be accurately determined (here it is 0.007502 seconds) and from this, the number of periods per second can be calculated as the reciprocal of the duration of the period (here 1/0.007502 = 133.298 periods per second). This measurement of *fundamental*

frequency, (f_0), was originally called cycles per second (cps) but in 1930 the unit of frequency was renamed *hertz* (Hz), in honour of the German physicist Heinrich Rudolf Hertz, who first conclusively proved the existence of electromagnetic waves.

The fundamental frequency of an utterance can be automatically displayed as a pitch curve, f_0 as a function of time, as in Figure 13.3. Here the pitch appears either as a continuous function when all sounds are voiced (cf 13.3a) or as a discontinuous function (cf.13.3b), when the utterance contains voiceless phonemes, like /t/ and /p/, for which there is no measurable pitch. In fact, speakers do not necessarily hear a difference in pitch between the two utterances. This corresponds to the observation (Nooteboom, 1997) that we do not perceive the observable discontinuities of raw pitch-patterns unless they are longer than about 200 ms: human perception appears to unconsciously bridge the silent gap by filling in the missing part of the pitch contour.



(a) (b)

Figure 13.3: Speech signal and f_0 for the French phrases: (a) *A ma maman* (to my Mummy) (b) *A ton papa* (to your Daddy) pronounced with a declarative intonation pattern.

13.2.2 Automatic models of pitch

Many automatic models of pitch have been developed, both for speech synthesis and for the empirical study of intonation patterns.

All these models, whether oriented towards the perception, the production or the acoustic realisation of pitch patterns, take as input the acoustic parameter of fundamental frequency (f_0), measured in hertz (Hz).

The search for an appropriate scale for measuring fundamental frequency was part of a systematic attempt, in particular by researchers from the 'Dutch school' (for a comprehensive summary of the work of this school, see 't Hart et al., 1990), to develop a model of the way in which pitch is perceived. This was done by stylising raw fundamental frequency patterns as a sequence of straight lines, such that when the stylised frequency is used to resynthesise the utterance,

the result is judged to be perceptually equivalent to the original intonation pattern.

Following work by House (1990) on the tonal perception of speech, Piet Mertens (Mertens & d'Allessandro, 1995; Mertens, 2004) developed an algorithm called *Prosogram* for the semi-automatic transcription of pitch, which assumes, somewhat controversially, that the perceptual segmentation of speech into syllables is prior to, and fundamental for, the perception of pitch. The following figure, from the Prosogram website (Mertens, 2018), illustrates the application of the stylisation algorithm to an utterance in French.

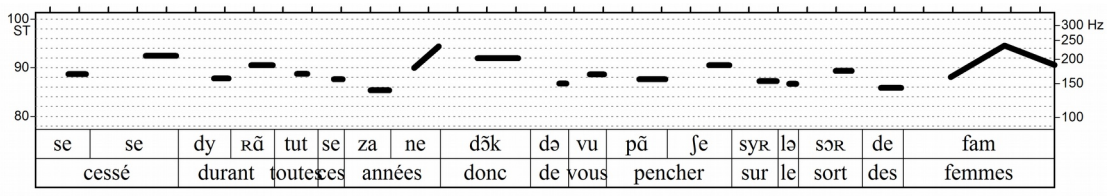


Figure 13.4 The output of the *Prosogram* algorithm applied to the French utterance "cessé durant toutes ces années donc de vous pencher sur le sort des femmes" (stopped during all these years then looking into the fate of women).

Another approach has been to attempt to model the way in which pitch is produced by speakers. In particular, work by Fujisaki and his colleagues has applied a model of pitch production (Fujisaki & Nagashima, 1969; Fujisaki, 2004) to a large number of languages, including several tone-languages, analysing an intonation pattern as the superposition of a sequence of phrasal components and of shorter accent components. These components are added in the logarithmic domain to produce a raw fundamental frequency curve as illustrated in Figure 13.5 (from Fujisaki, 2004).

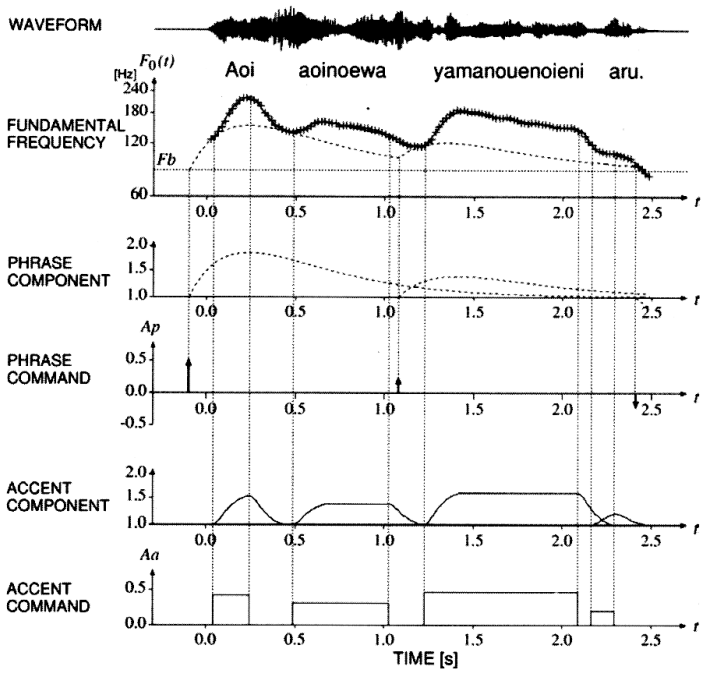


Figure 13.5 Analysis-by-synthesis of the Japanese sentence "Aoi aoinoewa yamanouenoieni aru." (The picture of the blue hollyhock is in a house on top of the hill) as a superposition of phrase components and accent components. From Fujisaki 2004.

A third approach has been to model directly the acoustic data, i.e. the f_0 curve. Fitting a raw f_0 curve with a mathematical model is not a simple straightforward problem due to the fact that fundamental frequency curves are not always continuous: unvoiced portions of the utterance have no associated f_0. Even when the curve is continuous, it is often not smooth and this type of irregularity is hard to model simply.

The discontinuity and irregularity of the f_0 curve is generally due to the presence of obstruents in the utterance, stops and fricatives, which either interrupt the curve (for voiceless obstruents) or make it irregular (for voiced obstruents). The effect of these consonants has been called *micromelodic* as distinct from the *macromelodic* characteristics of larger pitch movements associated with accents and intonation patterns (Di Cristo & Hirst, 1986). Micromelodic effects can be seen as a subset of more general *microprosodic* effects, specifically related to the local variability of the f_0 curve.

Micromelodic effects, then, are caused by the aerodynamic characteristics of the articulation of different phones. Phones like vowels and sonorants, which hardly obstruct the airflow, have virtually no micromelodic effect whereas stops and constrictives disturb or interrupt the flow of air through the vocal tract.

Linguists have known for a long time that fundamental frequency curves obtained from utterances containing only sonorants and vowels are much better

behaved than raw f_0 curves obtained from unrestricted speech. It is for this reason that linguists have often constructed sentences consisting of mainly sonorants and vowels such as Eva Gårding's *Madame Marianne Mallarmé har en mandolin från Madrid* (Madam Marianne Mallarmé has a mandolin from Madrid) for Swedish (Gårding, 1998), Annti Iivonen's *Laina lainaa Lainalla lainen* (Laina lends Laina a loan) for Finnish (Iivonen, 1998) or, the example from Figure 13.5, Hiroya Fujisaki's *Aoi aoinoewa yamanouenoieni aru* (The picture of the blue hollyhock is in a house on top of the hill) for Japanese (Fujisaki, 2004).

A raw intonation pattern, then, can be interpreted as the interaction between two independent components: a macromelodic component determined by the accentuation and intonation of the utterance and a micromelodic component determined by the segmental phonemes. If we compare two simple utterances in French like *A ton papa* (to your daddy) and *A ma maman* (to my mummy), pronounced with a declarative intonation pattern, we can see that there is the same underlying macromelodic pattern for the two utterances and that the surface differences are simply due to the different phonemes of the utterances, voiceless stops in Figure 13.3a and sonorant nasals in Figure 13.3b.

What is particularly worth noting is that the f_0 curve shown in Figure 13.3a is practically superposable on that of Figure 13.3b. It seems as if the f_0 curve continues to change during the voiceless segments of the utterance even though this is not, of course, visible. This is not as surprising as it may at first seem if we think in terms of a continuous change of the tension of the vocal folds, which can, of course, continue to change even during voiceless segments.

Notice in particular that the rise and fall on the two syllables of *papa* do not begin at the onset of the vowels: the f_0 at the vowel onset is already considerably different from that at the end of the preceding vowel. This idea of a continuously varying underlying pitch contour is not the model which is generally assumed in phonological descriptions of tonal and intonation contours. In the majority of these studies, it is assumed that tones are directly associated with vowels (cf. Halle & Vergnaud, 1987 pp 4-5; Goldsmith, 1990 p. 44 for examples) and that the fundamental frequency observed on the consonants is simply an interpolation between the tones on the vowels.

The fact that the f_0 curve follows the same trajectory in utterances with voiceless consonants as the smooth and continuous curve observed on the utterances with sonorants, however, and in particular the fact that the curve continues to evolve during the non-voiced portions of the utterance, seems fairly convincing evidence that the planning of these curves is the result of an

underlying macromelodic pattern on which the micromelodic variations are superposed.

The macromelodic component of an intonation pattern, then, has, we can assume, the two characteristics of being smooth and continuous. This is fortunate because, as mentioned above, modelling a discontinuous or irregular function is much more difficult than one which is continuous and smooth.

Once we have a macromelodic profile, we can derive the micromelodic profile by dividing each value of the raw f_0 curve by the corresponding value of the modelling function. Such a modelling technique is not simply a stylisation of the f_0 curve: the raw curve has actually been factored into two orthogonal components without any loss of information. For speech synthesis it is of course possible to model the micromelodic profile itself and to use this to improve the segmental quality of the utterance (for an application to Arabic see Chentir et al., 2009). For the study of intonation, the resynthesis of the utterance with the macromelodic profile is generally of sufficiently high quality to study the nature of the underlying intonation pattern.

One of the simplest ways to model a smooth continuous function like that of Figure 13.3a is as a piecewise sequence of transitions between successive points on the curve. We can call these points *anchor points.* In previous work (e.g. Hirst, 2007) these points were referred to as *target points*. The term *anchor points* is probably more appropriate, since the anchors do not necessarily have any specific psychological reality for the speaker and listener. The advantage of a piecewise function over a global function is that each segment of the curve is defined locally by its own set of parameters, which means that a modification of one portion of the curve does not entail modifications throughout the rest of the curve. The simplest model, of course, would be a linear transition between two anchor points, as was used in the perceptual model of the Dutch school, mentioned above ('t Hart et al., 1990).

Naturally occurring f_0 curves, of course, are not linear but curvilinear. A number of mathematical functions have been used in the past to model such functions. One of the simplest of these is a quadratic transition, corresponding to a constant acceleration followed by a constant deceleration of the pitch change. A continuous piecewise quadratic function is known as a quadratic *spline* function and has been in use since the 1980s to model intonation patterns using an algorithm called *Momel* (for "modelling melody") (Hirst, 1981; Hirst & Espesser, 1993).

The Momel model is in fact formally equivalent to a subset of the contours which can be produced by the Rise/Fall/Connection (RFC) model of intonation later developed by Paul Taylor (1995) as a tool for speech synthesis. The only

difference is that the RFC model allows linear interpolations between two successive anchor points as well as quadratic interpolations. In fact, if two successive anchor points have the same value of f_0, then the transition will be linear (i.e. flat) with Momel too. It is, naturally, an empirical question whether there exist cases where a non-flat linear transition gives a better approximation to an f_0 curve than a quadratic one.

The original implementation of Momel allowed the user to define anchor points manually by clicking on a representation of the f_0 curve on the computer screen. The user could then resynthesise the utterance using PSOLA resynthesis. This can be done today with the software package Praat (Boersma & Weenink, 2019) by creating a Manipulation object and removing and adding Pitch points manually. Praat displays the Pitch curve with linear interpolation between the Pitch points but an implementation of the quadratic spline function can be obtained by the command **Interpolate quadratically...**.

Manual modelling of f_0 is, of course, highly subjective and it was for this reason that an automatic version of the algorithm was developed, based on the experience of using the manual implementation of the model over a period of several years. The algorithm, which is described in detail in Hirst et al. (2000) uses a form of robust regression to optimise the modeling of raw fundamental frequency curves with a quadratic spline function.

The algorithm was later evaluated on a corpus of read speech in 5 languages (corpus Eurom1) during the course of the Multext European project (Véronis et al., 1994). Examination of the errors in the f_0 modelling showed that one type of error in particular occurred systematically. This concerned a pitch rise before a silent pause where, frequently, the algorithm missed the final pitch of the rise entirely.

The Momel algorithm has since been implemented as a Praat plugin (see Hirst, 2007), which makes it possible to use its functions directly from the Praat menus without needing to manipulate scripts. The systematic error observed previously was corrected by a special treatment before silent pauses, extrapolating the final rise to estimate the closest anchor point that will produce such a rise. An evaluation of the improved algorithm was carried out on a corpus of read speech in Korean (Hirst et al., 2007). It showed a significant and systematic improvement on the fitting of the modelled curve to the raw fundamental frequency as compared to the older version of the algorithm.

It is, naturally, desirable that the modelling tools we use should be as theory-neutral as possible. Complete neutrality, though, is obviously not entirely feasible, since any model necessarily makes some assumptions about the nature of underlying representations, as we saw above in the discussion of whether the

underlying contour should be based only on the contours observed on the vowels or whether it should be modelled as a continuous underlying contour.

If not theory-neutral, the Momel algorithm might be described as *theory-friendly*. That is the algorithm can be compatible with a number of different theoretical approaches to the description of speech melody. It has, in fact, been used in the past as a first step for modelling with the Fujisaki model (Mixdorff, 1999). It has also been used as first step for ToBI for both English (Maghbouleh, 1998; Wightman & Campbell, 1995) and Korean (K-ToBI) (Cho & Rauzy, 2008).

13.2.3 Pitch scales

The output of pitch detection algorithms normally uses a linear scale of Hz. Although there has been some controversy on this point, it is generally accepted that the perception of pitch differences is not linear – that is a pitch rise or fall of, say, 50 Hz is perceived as being smaller for a higher-pitched voice (such as that of a woman or a child) than for a lower pitched voice (such as that of a man). Similarly, intonationally equivalent utterances, when produced by a male and a female speaker, may sound the same to the listener, i.e. they may convey the same linguistic or paralinguistic functions, even though, when measured in Hz, they have different sizes: the pitch movements measured in hertz will generally be larger for a female voice, as her pitch range is on average higher and wider than that of a male voice (Graddol, 1986; Hermes & van Gestel, 1991).

The linear hertz scale should not, consequently, be used to measure differences in frequencies, e.g. when analysing the span of a speaker's pitch range or the pitch movements of a melodic contour. This is why, in intonation research, the hertz scale is generally transformed to a logarithmic scale (e.g. semitones or octaves) or to a psychoacoustic scale (e.g. mel, Bark or ERB) as described below.

In order to take into account the non-linear nature of pitch perception, many studies of intonation adopted the solution of using a musical scale. As early as the 18th century, Joshua Steele (1778) used a bass viol to imitate the melody of speech and transcribed the pitch using a very detailed system of transcription based on musical notation.

A musical scale effectively converts the values to a logarithmic scale. Many studies (e.g. Jassem, 1952, 1971; 't Hart et al., 1990; Fant, 1968, 2004) have used a scale in equal tempered semi-tones to represent pitch intervals, by means of a mathematical formula like:

(1) interval = $12*\log_2(h1/h2)$

where h1 and h2 are the two limits of the interval.

A frequency in Hz can be represented as a musical note using the following formulas, where *reference* is the reference pitch, usually 440 Hz for standard concert pitch (= A4, the A above middle C).

(2) a. semitones = $\log_2(\text{frequency}/\text{reference})*12+57$
b. octave = round(semitones / 12)
c. note = round(semitones) mod 12 + 1
d. error = semitones - round(semitones)

For a frequency such as 157 Hz, this gives octave = 3, note = 4, error = +0.159. Using the list of note names {C, C#, D, E ♭ , E, F, F#, G, A ♭ , A, B ♭ , B}, we can then identify the frequency as corresponding to E ♭ 3 with an error of +0.159 semitones. These formulas have been implemented as a Praat script *diapason.praat* (Hirst 2012) which can also convert a musical note such as E ♭ 3 to a frequency (= 155.56 Hz).

Log scales are used as an approximation to the perceptual value of pitch height, following the Weber-Fechner law (https://en.wikipedia.org/wiki/Weber-Fechner_law), which states that the magnitude of a perceived sensation is directly proportional to the logarithm of the physical magnitude of a stimulus. Psychoacoustic pitch scales, have however been claimed to be closer to the specific perception of the pitch of speech sounds.

Several studies have suggested that the optimal scale for pitch intervals is intermediate between a linear scale and a logarithmic scale using the so-called psychoacoustic scales measured in *mels*, *Barks* or *ERBs*. Psychoacoustic scales aim to model the way spectral information is processed in the human auditory system. They were designed to be as optimal as possible to measure pitch intervals. They provide steps which correspond to pitch intervals that are perceived to be of equal size.

The *mel scale* (Stevens, Volkman and Newman, 1937) is a perceptual scale, which was based on listeners' subjective judgements of equal pitch magnitude using sinusoid tones to divide frequency ranges into sections. According to Beranek (1949):

> Mel is the unit of pitch. It is so designed that a 1000-cycle tone 40 dB above threshold has a pitch of 1000 mels. [p28]

There is no single accepted formula for converting Hz to mels. O'Shaughnessy (1984) gives the formula:

(3) $mel = 1127 * \ln(1 + f/700)$

All the following formulas are given using natural logarithms *ln*(f), although some of the original versions of the formulas used logarithms with base 10: *log10*(f). The value of *ln*(f) is the same as *log10*(f)/*log10*(e) = *log10*(f) * 2.302585.

Fant (2004) gives a formula for what he calls the *technical mel scale*:

(4) $mel = 1000*(\ln(1 + f/1000)/\ln(2))$

while Praat (Boersma & Weenink 2019) uses the formula:

(5) $mel = 550 * \ln(1+f/550)$

which, unlike most other versions, does not give 1000 mels for 1000 Hz.

The *Bark scale* **(**Zwicker 1961) was defined so that each critical band of human hearing has a width of one *Bark*. There are many formulae that exist to convert frequency values in *hertz* to *Bark* values, such as Traunmüller's approximation (1990), using the formula:

(6) $Bark = 26.81f/(1960+f) – 0.53,$

Boersma & Weenink (1992) and Fant (2004) give:

(7) $Bark = 7 \ln(f/650 + \sqrt{(1+ (f/650)^2)})$

where *f* is the frequency in Hertz.

The *ERB scale* (Equivalent Rectangular Bandwith), like the Bark scale, was defined to be closely related to the critical bandwidth and was measured from the ability to detect sinusoids in the presence of noise.

Moore & Glasberg (1983) give the following formula:

(9) $ERB = 6.23* f^2 + 93.39 * f +28.52$

then in 1996 they give:

(10) $ERB = 24.7 * (4.37 f / 1000 + 1)$

Hermes & van Gestel (1991) use the formula:

(11) $ERB = 7.253 * \ln(1 + f/165.4)$

while Boersma & Weenink (1992) use:

(12) $ERB = 11.17 * \ln((f +312) / (f + 14680)) + 43$

Several experimental studies have shown that the logarithmic and psychoacoustic scales account better for listeners' perception of pitch differences than a linear scale. There is, however, no consensus as to which scale (or formula) is preferable and for what tasks.

It has often been claimed that these psychoacoustic scales are linear for lower frequencies, under 500 or 1000 Hz and logarithmic for higher frequencies. Umesh et al. (1999), however, tested a large number of formulas to fit the data from the original presentation of the *mel* scale (Stevens et al. 1937). For the region below 1000 Hz, the best fit was given by:

(13) $mel = 3294 - 3080*\ln(f) + 773*(\ln(f))^2$

although several other mathematical functions also gave a good fit.

They conclude that:

> there is no evidence that there are two qualitatively different regions. In particular there is no evidence that the lower region is linear and the upper region is logarithmic. (p 220)

Figure 13.6 shows, from left to right (or top to bottom), the log scale, the ERB scale (Hermes & van Gestel 1991), the Bark scale (Fant 2004), the me scale (O'Shaughnessy 1987) and the linear scale.Each scale is normalised on the y-axis between fn(50) and fn(500), where fn is the corresponding function. It can be clearly seen from this figure that all the psychacoustic scales are between linear and logarithmic.

Figure 13.6 From left to right (or top to bottom) the log scale, the ERB scale], the bark scale, the mel scale and the linear scale. Each scale is normalised on the y-axis between *fn*(50) and *fn*(500), where *fn* is the corresponding function.

Traunmüller (1997) gives a good description of the auditory processes which are behind the perception of pitch. After an account of the different auditory scales, he concludes:

> In order to visualize pitch contours in speech, it is suggested to use a semitone scale or to scale frequency (or period) logarithmically.

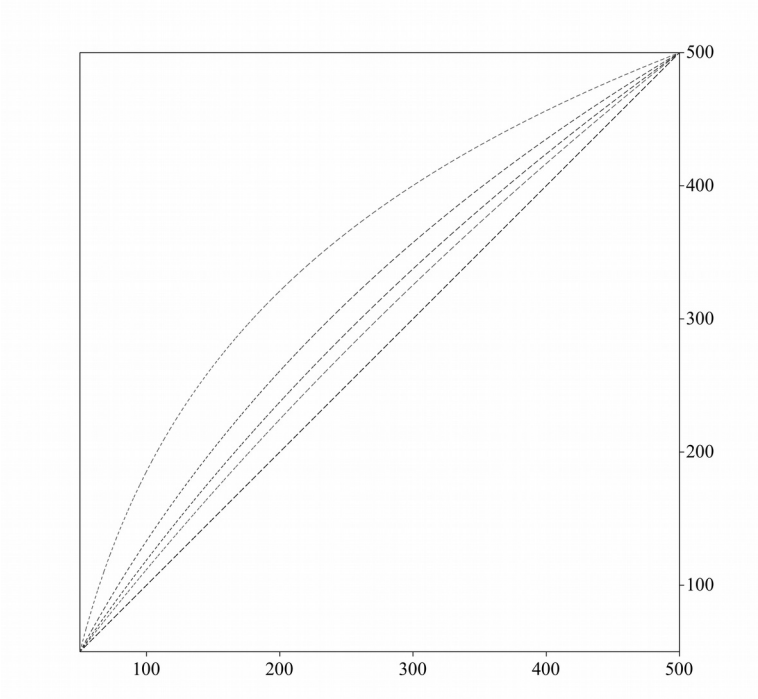


Figure 13.6 From left to right (or top to bottom) the log scale, the ERB scale], the bark scale, the mel scale and the linear scale. Each scale is normalised on the y-axis between *fn*(50) and *fn*(500), where *fn* is the corresponding function.

In a production experiment (Nolan 2003), subjects were asked to replicate the pitch contours of utterances produced by a female and a male speaker in their own voice. In order to evaluate which scale best accounts for a listener's perception of intonation equivalence, the differences between the pitch span of each template and that of its replication were calculated, and compared using the *hertz*, *semitone*, *ERB*, bark and *mel* scales. Smaller differences were found for the semitone and ERB scales, suggesting that the optimal scale for comparing pitch contours is logarithmic or nearly logarithmic.

13.3 Critical Issues

One of the most critical problems in the study or modelling of pitch is to obtain an accurate estimate of the fundamental frequency.

13.3.1 Direct observation of vocal folds

The most accurate measurement of fundamental frequency can be obtained by directly examining the movements of the vocal folds during speech production, or their associated muscle activation. A number of techniques such as *laryngeal electromyography*, *laryngoscope*, and *depth kymography* have been used for this task. Less invasively, an *electroglottograph* (also known as *laryngograph*) can be used to measure the electrical impedance through the neck at the level of the larynx, giving a direct image of the opening and closing of the vocal folds (Fourcin & Abberton 1971).

13.3.2 Detection from the acoustic signal

Detecting fundamental frequency from the acoustic signal is more complicated since the semi-periodic glottal wave-form is distorted by its passage through the vocal tract.

A large number of different algorithms have been proposed to estimate the periodicity of the signal, and hence the fundamental frequency, calculated as the reciprocal of the duration of the period. For a very thorough background to pitch detection algorithms, see Hess (1983) who notes:

> For a number of reasons (…) the task of pitch determination has to be counted among the most difficult problems in speech analysis' (p vii).

The Wikipedia page *Pitch Detection Algorithm* (Wikipedia, 2018) gives a useful update and references for more recent algorithms. Essentially, pitch detection algorithms are of two basic types: time-domain approaches and frequency domain approaches.There are also algorithms which use a combination of both approaches.

The time-domain approach looks for semi-periodicity in the acoustic waveform by comparing two consecutive short portions of the signal, the duration of which corresponds to the shortest period considered acceptable. The duration of the window is then progressively increased up to that of the longest period considered acceptable. The window size giving the best match is taken as the duration of the period at that point. Algorithms of this type such as ACF autocorrelation function), AMDF (average magnitude difference function), YIN (from oriental 'yin' and 'yang'), MPM (McLeod Pitch Method) generally use a form of autocorrelation to compare the signal in consecutive windows.

The frequency-domain approach works by creating an estimate of the frequency spectrum and then looking for the best candidate for a harmonic

interval on the spectrum which is then taken as the fundamental frequency. Algorithms of this type such as HPS (Harmonic Product Spectrum), cepstral analysis, maximum likelihood, spectral comb function(Martin 1981, have the advantage that they can provide a more reliable estimate of fundamental frequency with degraded speech signals such as telephone speech.

In some recent algorithms,e.g. YAAPT (Yet Another Algorithm for Pitch Detection), a combination of the time-domain and frequency-domain approaches is used and the final pitch is computed by applying dynamic programming to the candidates from the two approaches.

For details and references concerning most of these algorithms, see Wikipedia (2018).

13.3.3 Pitch tracking errors

While current pitch detection algorithms perform relatively well, they may still result in pitch tracking/detection errors. Vocal fold cycle irregularities (e.g. creaky voice), rapid changes in f_0 and noisy environments may all be sources of aberrant values in the f_0 detected (Kiessling et al 1995; Brønsted, 1997). Most common errors come from the fact that the algorithm fails to accurately estimate a frequency value or determine periodicity vs. non-periodicity for speech segments. *Octave errors* are common examples of defective f_0 detection, where the estimated frequency is half or the double that of the perceived value.

Other types of f_0 perturbations are due to the intrinsic nature of phones and to their co-articulation (House & Fairbanks 1953; Di Cristo & Hirst 1986; Silverman 1986; Hanson 2009). Pitch skips at the onset of vowels are examples of such perturbations, resulting from the aerodynamic characteristics of the articulation of phones like stops and constrictives.

Many errors of pitch tracking can fortunately be avoided by an appropriate choice of pitch settings as described below.

13.3.4 Estimating optimal values for pitch floor and ceiling

Most speech analysis programs use default parameters defining the minimum and maximum values which are allowed for the f_0. In Praat, for example, these values, referred to as *Pitch Floor* and *Pitch Ceiling*, are set by default to 75 and 600 Hz respectively. Unfortunately, these default parameters are rarely satisfactory. In order to reduce f_0 tracking errors, the authors recommend in their manual (available in the Praat software and on the Praat website http://www.praat.org) to set these parameters to an estimate of the speaker's pitch range. They suggest the values 100-500 Hz for female speakers and 75-300 Hz for male speakers.

The choice of parameters can also be automated by using a two-pass detection process (De Looze & Hirst 2008). In the first pass, extreme values (e.g. 60 and 700) are used as *pitch floor* and *pitch ceiling*. From the f_0 thus obtained, the first and third quartiles (*q1*, *q3*) of the pitch distribution are then calculated. These are rather robust with respect to pitch detection errors, which are usually located in the upper and lower percentiles of the distribution (De Looze, 2010). A value for the pitch floor can then be calculated as 0.75 *q1*. This has been shown empirically to provide a fairly optimal estimate of the pitch floor independently of the speaker's actual pitch range (De Looze, 2010). The pitch ceiling tends to be more variable, depending on the degree of expressivity of the speech. For fairly non-emphatic speech, the ceiling can be fixed to 1.5 *q3*, for more emphatic speech the ceiling would need to be raised to something like 2.5 *q3*.

13.3.5 The difference between pitch and f0

Even with all the precautions we have described, there are sometimes cases where the output of the pitch detection algorithm does not correspond to what is perceived. Mark Liberman on his website *The Language Log* (Liberman, 2017) gives an example of a recording of the phrase: “once the eggs hatch” which most people hear as containing a rise from the syllable “eggs” to the end of the phrase. As we can see in Figure 13.7, however, the detected f_0, even using the optimised maximum and minimum settings, described in Section 13.3.4, shows a significant drop to the final syllable. If we set the minimum f_0 higher, say to 125 Hz, then no f_0 is detected at all on the main part of that syllable.

In conclusion, as Mark Liberman comments:

> (the psychological dimension of) *pitch* is not the same as the (physical dimension of) *fundamental frequency*.

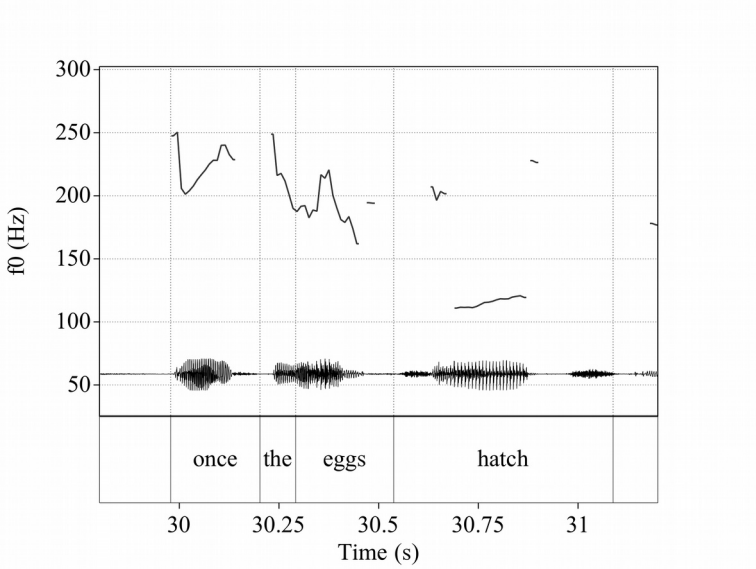


Figure 13.7 Fundamental frequency of the phrase *once the eggs hatch*. (from Liberman 2017).

13.4 Recent Research

13.4.1 Semitones and octaves

The semitone has frequently been used as the basic unit for a logarithmic scale for the analysis and display of the pitch of speech, due partly to the fact that a semitone is approximately the minimum interval that normal listeners (without special musical training) can distinguish (Jassem 1952 p 37, citing Zwirner & Zwirner 1937) and also to the fact that the equal tempered semitone is the basic interval for numerous Western musical instruments (notes on a piano, frets on a guitar). For more accurate measurements, the semitone can be divided into 100 cents. This is used in particular for the comparison of similar intervals in different tuning systems.

In a recent study, however, (De Looze & Hirst, 2014) we suggested that the octave, rather than the semitone, is the most natural interval for analysing speech.

Following evidence reported in several studies based in neuronatomy, neurophysiology, behavioural studies, speech production as well as speech perception (see De Looze & Hirst 2014 and below), we recommended the systematic use of the octave (o) and its subdivision the millioctave (mo) for the study of pitch. The mo gives approximately the same degree of precision as the cent (1 mo = 1.2 cents) and has the advantage of being in conformity with the general practice of the International System of Units: SI, in which prefixes corresponding to an exponent divisible by 3 (e.g. n, m, k, M) are generally preferred.

As a derived SI unit, the octave can be defined as:

(14) $o = \log_2(p^{-1})$

where *p* is the duration in seconds of a period.

The semitone is, in fact, the product of a complex history of Western classical music culture, corresponding to the division of the octave into 12 equal intervals. This idea was first described in a treatise published in China in 1584 (Kuttner, 1975). In Europe, the scale of 12 equal semitones (referred to as equal temperament) has been used increasingly, particularly in the last century, to tune keyboards, replacing the natural scale ('just intonation') previously used, or Bach's well-tempered scale (Lindley 2001). All these scales were the result of a search for a compromise, which would allow musicians to modulate from one scale to another without introducing major discord and without having to switch keyboards.

In different civilizations at different times, musical scales have in common the fact that the names of the notes can be repeated indefinitely within the physical limits of sound production. This circularity (also known as *chromatic repetition*) appears, in fact, to be universal, and seems to stem from a physiological basis of human perception (Braun & Chaloupa, 2005; Braun 2006) including that of neonates (Liu et al, 2009) and also that of rhesus monkeys (Wright et al, 2000). It was observed as early as the 1960s, in an anatomical study of a cat, that the auditory thalamus is organised in stacked layers or laminae. It was suggested that this organisation may have a specific function in the processing of acoustic frequencies (Morest, 1965). Morel (1980) and Imig and Morel (1985) later demonstrated that the auditory thalamus of the cat actually contains a neural chroma map, underlying an octave architecture, where octaves are represented by clusters of neural laminae. While the functional role of the mammalian auditory thalamus octave topography still needs to be determined, recent research by Braun and Chaloupka (2005) has suggested that it may cause, as a side effect, the octave circularity of pitch that has been observed in the rhesus monkey as well as in humans. Their study investigated the effect on a musician with absolute pitch, of the neurotropic medical drug carbamazepine (CBZ), known to have a down-shift pitch side effect, in order to better understand the mechanism of octave circularity of pitch. They observed in their subject, during a pitch identification task, an internal tone-scale or chroma representation. When CBZ was taken, a pitch shift was indeed observed but the pattern of tone representation remained unchanged. This suggests that the human brain may be hard-wired for octave-circular pitch perception. In any case, it is the octave, not the semi-tone, which appears clearly

as the basic unit for the natural perception of the pitch of speech sounds and music.

The first author has suggested elsewhere (Hirst, 1981, 1983) that there may also be a physiological explanation for the octave and half octave as a basis for the production of melodic intervals. Hirst (1983) reported an experiment where these two intervals (octave and half-octave) were observed as modal values in a task of producing varied contours on isolated syllables in French, *oui* and *non*.

If we assume that the vocal folds behave like vibrating strings, the relationship between tension and frequency is governed by Mersenne's law (https://en.wikipedia.org/wiki/Mersenne's_laws), which states that the frequency of a vibrating string is proportional to the square root of its tension. A doubling of the tension would consequently correspond to a rise of half an octave. This might explain why the intervals – octave and half octave (respectively 12 and 6 semitones) – have been found to be so frequent in the production of speech melody, even though a rise or fall of a complete octave on a single syllable is certainly not perceived in its entirety (Rossi, 1971).

The use of the semi-tone has paradoxically had the negative effect of masking the importance of the octave as a basic unit in pitch production and perception. A number of studies on pitch range have reported an interval close to an octave (= 12 sts) or half octave (= 6 sts) without drawing attention to this fact, or perhaps, even, sometimes without having noticed it.

't Hart et al. (1990) note that:

> In Dutch intonation, excursions most frequently vary around six semitones (…). In German intonation, the excursion (for full-size movements) can be taken as ten semitones, only slightly less than in British English intonation. (p 53)

Paesche & Sendlmeier (2000) reported an f_0 mean at the beginning of sentences produced in neutral, happy, angry and scared voices of 6.72, 12.64, 12.52 and 12.38 sts respectively. If we calculate the difference between the mean f_0 of neutral voice and that of the other types of voice, we find for each 'arousal' voice a shift of half an octave.

The intervals octave and half-octave may play a specific role in speech production. Braun (2001) investigated the pitch contours of utterances produced under two conditions (in a normal voice in a quiet room vs. in a louder voice when exposed to noise over headphones), and observed a raising of half an octave for the increased loudness condition. A rise of a half an octave or an octave may be used to convey specific linguistic and paralinguistic functions in

speech, e.g. signalling focus, topic change, turn-taking as well as expressing arousal.

13.4.2 The Octave-Median scale

In De Looze & Hirst (2014), we further recommended the use of the Octave-Median (OMe) scale for the display of the fundamental frequency curve and for comparison of utterances produced by speakers with different pitch ranges. The value on the OMe scale can be calculated as:

(15) $ome = \log_2(Hz/median)$,

where *Hz* is the value of the frequency in hertz and *median*, the median, also in hertz, of the speaker's fundamental frequency.

In this transform, the reference is given by the median of the speaker's pitch range.

It should be noted that the median is a far more robust measurement of the central tendancy of a pitch distirubution than the mean. Unlike the mean, the median is generally not affected by pitch errors at the top or bottom of the distribution. The median is also independent of the pitch scale so that the same value will be selected for a linear scale as for a logarithmic (semitone or octave) scale.

The OMe scale was defined following an analysis of several corpora of neutral non-emphatic speech in French and in English (De Looze, 2010), where it was found that the speakers' non-emphatic pitch range tends to lie within one octave around the median pitch (i.e. -0.5 : +0.5 on the OMe scale). The bottom of the central octave of the speaker's voice is then half an octave below the median, while the top is half an octave above.

Using the Momel-INTSINT algorithms (Hirst, 2007), we investigated the relationship between the median, minimum and maximum values of a speaker's pitch range. The INTSINT algorithm uses a symbolic alphabet to code the anchor points found by the Momel algorithm. These tonal symbols T(op), B(ottom), M(id), H(igher), L(ower), S(ame), U(pstepped), and D(ownstepped), can be used to generate a synthetic intonation pattern from two parameters representing the mid point (key) and the span of the speaker's pitch range.

We found, in fact, very strong correlations between the speakers' median pitch and their bottom (B) and top (T) values.

(16) a. $B= 0.706*median$ ($R^2=0.92$)
b. $T=1.561 * median$ ($R^2=0.91$)

In Figure 13.8 the corresponding linear regressions are plotted in solid lines and dashed lines represent the intervals (from top to bottom) + octave, + half-octave, unison, - half-octave and - octave with respect to the median. The linear regression on the mean of the bottom tones (B) coincides almost exactly with the half-octave below the median so that the two lines are not distinguishable in the figure. That of the average of the top tones (T) falls between half an octave and one octave above the median.

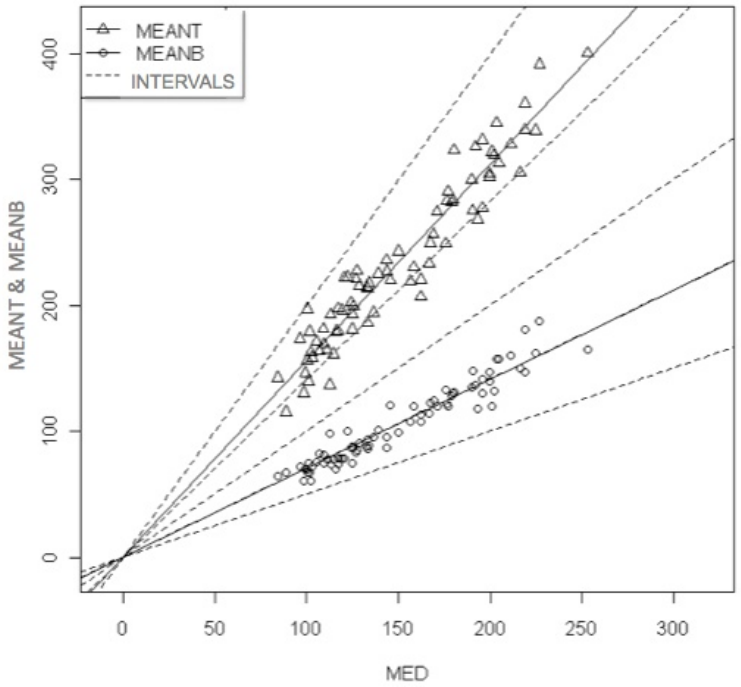


Figure 13.8 Graphical representation of the average bottom tones (B) and the average top tones (T) using the Momel-INTSINT algorithms compared to the speaker's median pitch (from De Looze & Hirst 2008).

The coefficient 0.706 corresponds almost exactly to half an octave (log2(0.706) = -0.502) and the coefficient 1.561 is just slightly over half an octave (log2(1.561) = 0.642). These results suggested that the average of the high tones and the average low tones, i.e. the limits of the range of a speaker, for unemphatic speech, usually correspond to about one octave, centred on the speaker's median. This also means that it is possible, at least as a reasonable approximation, to predict the limits of the register of a speaker and hence its span, from the median of the distribution of f0.

Within the frame of an OMe representation, top lines of the display should not be thought of as physical obstacles for speakers. Rather, in more spontaneous corpora, larger pitch ranges - up to two octaves are likely to be expected. Pitch often goes beyond these lines (De Looze 2010), particularly in the case of the top line, but when it does so, it may be taken as a good sign that the speech is expressive or signalling important information.

13.4.3 The graphic representation of pitch patterns

Figure 13.9 illustrates the sentence "What can I have for dinner tonight?" read by one female and one male speaker. The visualization of these recordings was obtained automatically from the signal and TextGrid using the Praat plugin ProZed (Hirst, 2015).

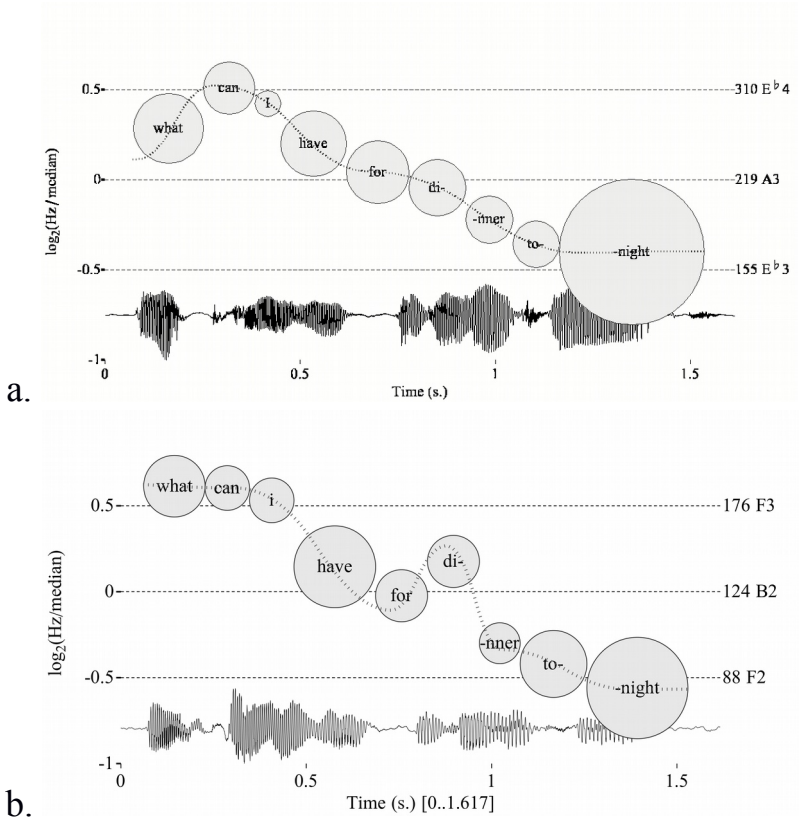


Figure 13.9 Pitch patterns for 2 speakers, 1 female (top) and 1 male (bottom) reading the sentence What can I have for dinner tonight? (see text).

The diameters of the circles correspond to the syllable durations and the dotted line corresponds to the Momel curve. The horizontal dashed lines correspond to the speaker's median (middle line) and a half octave above and below the median, delimiting the speakers unemphatic pitch range

corresponding to the median-centred octave. The values of the Median and the Top and Bottom of the central octave are given in Hz and as musical notes. With this technique, the optimal parameters for the analysis of the fundamental frequency of the speaker can be automatically determined from the median pitch.

13.5 Best Practice for Teaching and Learning

Students beginning to learn acoustic analysis are confronted with the fact that the visual representations of speech which they can obtain on their computers are an unfamiliar way of representing a very familiar object. Considerable experience in working with these representations is needed before being able to rely on intuitions about how the visual form of a waveform or a spectrogram relates to the corresponding sound.

In the specific case of pitch, the situation is slightly different. Nearly every student is more or less familiar with the standard classical notation of music even though skills in interpreting these representations vary greatly from individual to individual. In this notation, pitch height is represented by the vertical position of the note on the stave, so that it is fairly straightforward to generalise from this to using continuous lines which rise and fall in imitation of the pitch of a voice.

The main problem with interpreting raw fundamental frequency patterns comes from the fact that, as we saw in Section 13.2.1, these patterns are a mixture of continuous and discontinuous lines; it is not evident for a beginner to realise that Figures 13.3a and 13.3b, for example, ("A ma maman" and "A ton papa") represent the same intonation pattern with different segmental material. The solution to this is simple: a lot of practice. Students need to devote several hours to listening to spoken material and examining the corresponding pitch patterns. An efficient way to do this is to label by hand a corpus of speech and then to compare the labelling with the output of an automatic labelling system (e.g. Bigi, 2015), checking the pitch at the same time. Accompanying this with a stylised representation of the intonation such as those described in this chapter is also a very useful technique.

Attention should be drawn to the possibilities of pitch tracking errors as described in Section 13.3.3, in particular those associated with creaky voice or octave errors, the latter often due to an inappropriate choice of pitch floor and pitch ceiling.

It is strongly recommended, in case of doubt, to check the pitch measurements perceptually. With Praat (Boersma and Weenink, 1992), this can be done very easily by selecting the Pitch object and calling the command **Play**

pulses or **Hum**. It should be noted that to do this, the Pitch needs to be extracted as a separate object. It cannot be done using the display of Pitch in the sound editor. Many of the pitch tracking errors we have described can easily be identified as errors using this technique.

13.6 Future Directions

A number of unsolved problems could be the object of future research. Although many studies have been devoted to the perception of pitch there is still some uncertainty, as we mentioned in Section 13.2.3, about the optimal scale for representing the pitch of spoken utterances.

There is even less certainty about the relationship of pitch to basic physiological characteristics such as vocal tract length or the length of vocal folds of speakers. It seems obvious that speakers take such features into account when they perceive pitch, but current models of pitch detection do not make use of this type of information.

We suggested in Section 13.4.1 that the relation between pitch and the tension of vocal folds may follow Mersenne's law, which states that the frequency of a vibrating string is proportional to the square root of its tension. As far as we know there have been no empirical studies testing this hypothesis although it would be relatively easy to imagine ways to test this empirically.

A better knowledge of the way in which physiological constraints interact with perceptual constraints will certainly provide a solution to many of these questions.

13.7 References

Beranek, L. L. (1949). *Acoustical Measurements*. Acoustical Society of America. [revised edition 1988]

Bigi, B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician* (International Society of Phonetic Sciences), 111-112 (I-II), 54-69.

Boersma, P & Weenink, D. (2019). *Praat: doing phonetics by computer*. [computer program: Version 6.0.56, June 2019: http://www.praat.org]

Braun, M. (2001). Speech mirrors norm-tones: Absolute pitch as a normal but precognitive trait. *Acoustics Research Letters Online*, 2(3), 85-90.

Braun, M. (2006). A retrospective study of the spectral probability of spontaneous otoacoustic emissions: Rise of octave shifted second mode after infancy. *Hearing Research* 215, 39-46.

Braun, M. & Chaloupka, V. (2005). Carbamazepine induced pitch shift and octave space representation. *Hearing Research,* 210, 85-92.

Brøndsted, T. (1997). Intonation contours distorted by tone patterns of stress groups and word accent. In A. Botinis (ed) *Intonation : Theory, Models and Applications* (Proceedings of an ISCA workshop), Athens (Athanasopoulos), pp. 55-58.

Chentir, A., Guerti, M. & Hirst, D. J. (2009). Extraction of standard arabic micromelody. *Journal of Computer Science*, 5(2), 86–89.

Cho, H. & Rauzy, S. (2008). Phonetic pitch movements of accentual phrases in Korean read speech. In *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil.

De Looze, C. (2010). *Analyse et interprétation de l'empan temporel des variations prosodiques en français et en anglais.* PhD thesis, Université de Provence, Aix-en-Provence, France.

De Looze, C. & Hirst, D. J. (2008). Detecting changes in key and range for the automatic modelling and coding of intonation. In *Proceedings of 4th International Conference on Speech Prosody.* Campinas, Brazil. May 6-9, pp. 135-138.

De Looze, C. & Hirst, D. J. (2014). The OMe (Octave-Median) scale: a natural scale for speech melody. *Proceedings of the 7th International Conference on Speech Prosody*, Dublin. May 20-23, pp. 910-913.

Di Cristo, A., & Hirst, D. J. (1986). Modelling French micromelody: analysis and synthesis. *Phonetica*, 43 (1-3), 11-30.

Fant, G. (1968). Analysis and synthesis of speech processes. In B. Malmberg, ed., *Manual of Phonetics*. Amsterdam: North-Holland, pp. 173-177.

Fant, G. (2004). *Speech Acoustics and Phonetics*. Dordrecht: Kluwer.

Fourcin, A. J. & Abberton, E. (1971). First applications of a new laryngograph. *Medical and Biological Illustration,* 21, 172–82.

Fujisaki, H. (2004). Information, prosody, and modeling - with emphasis on tonal features of speech. In *Proceedings of the Second International Conference on Speech Prosody (Nara, Japan)*, pp. 1-10.

Fujisaki, H. & Nagashima, S. (1969). A model for the synthesis of pitch contours of connected speech. *Annual Report of the Engineering Research Institute*, 28, 53-60.

Gårding, E. (1998). Intonation in Swedish. In D.J. Hirst and A. Di Cristo, editors, *Intonation Systems. A Survey of Twenty Languages.* Cambridge: Cambridge University Press, pp. 117-136.

Goldsmith, J. A. (1990). *Autosegmental and Metrical Phonology*. Blackwell: Cambridge, Mass.

Graddol, D. (1986). Discourse specific pitch behaviour. In C. Johns Lewis, eds., *Intonation in Discourse*. Edinburgh: Croom Helmpp. 221-238.

Halle, M. & Vergnaud J.-R. (1987). *An Essay on Stress*. Cambridge, Mass.: MIT Press.

Hanson, H. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *Journal of the Acoustical Society of America,* 125, 425-441.

Hart ('t), J., Collier, R. & Cohen, A. (1990). *A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody*. Cambridge: Cambridge University Press.

Hermes, D. I. & van Gestel, I. E. (1991). The frequency scale of speech intonation. *Journal of the Acoustical Society of America*, 90, 97-102.

Hess, Wolfgang (1983). *Pitch Determination of Speech Signals. Algorithms and Devices.* Belin: Springer Verlag.

Hirst, D. J. (1981). Phonological implications of a production model of English intonation. *Phonologica*, 1980, 195-201.

Hirst, D. J. (1983). Structures and categories in prosodic representations. In Cutler & Ladd (eds). *Prosody: Models & Measurements.* Berlin: Springer, pp. 93-109.

Hirst, D. J. (2007). A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. In *Proceedings of the XVIth International Conference of Phonetic Sciences* (paper 1443), Saarbrücken, pp. 1233-1236.

Hirst, D. J. (2012). Diapason.praat [Praat script. Downloadable from: www.researchgate.net/publication/327764721_diapason].

Hirst, D. J. (2015). ProZed: A speech prosody editor for linguists, using analysis-by-synthesis. In Keikichi Hirose & Jianhua Tao, eds. *Speech Prosody in Speech Synthesis. Modeling and Generation of Prosody for High Quality and Flexible Speech Synthesis*. Berlin, Heidelberg: Springer Verlag, pp. 3-17.

Hirst, D. J., Cho, H., Kim, S. & Yu, H. (2007). Evaluating two versions of the Momel pitch modeling algorithm on a corpus of read speech in Korean. In *Proceedings of Interspeech*, volume VIII. Antwerp, Belgium, pp.1649-1652.

Hirst, D. J., Di Cristo, A. & Espesser, R. (2000). Levels of representation and levels of analysis for intonation. In M. Horne, ed., *Prosody: Theory and Experiment*. Dordrecht: Kluwer Academic Publishers, pp. 51-87.

Hirst, D. J. & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, 15, 75–85.

House, A., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, 25, 105-113.

House, D. (1990). *Tonal Perception in Speech*. Lund: Lund University Press.

Imig, T. J. & Morel, A. (1985). Tonotopic organization in ventral nucleus of medial geniculate body in the cat. *Journal of Neurophysiology* 53, 309-340.

Iivonen, A. (1998). Intonation in Finnish. In D.J. Hirst and A. Di Cristo, eds., *Intonation Systems. A Survey of Twenty Languages*. Cambridge: Cambridge University Press, pp. 331–347.

Jassem, W. (1952). *Intonation of Conversational English (educated Southern British)*. Wrocław: Wrocławskie Towarzystwo Naukowe. [PDF available from the Speech and Language Data Respository, at http://sldr.org/sldr000777/en]

Jassem, W. (1971). Pitch and compass of the speaking voice. *Journal of Phonetics*, 1, 59-68.

Jones, D. (1909). *Intonation Curves*. B.G. Leipzig; Berlin: Teubner.

Kiessling, A., Kompe, R., Niemann, H., Nöth, E., & Batliner, A. (1995). Voice source state as a source of information in speech recognition : Detection of laryngealizations. *Natoasi Series of Computer and Systems Sciences*, 147, 329-332.

Kuttner, F. A. (1975). Prince Chu Tsai-Yu's life and work: a re-evaluation of his contribution to equal temperament theory. *Ethnomusicology,* 19(2), 163-206.

Liberman, M. (2017). Pitch contour perception. [webpage: http://languagelog.ldc.upenn.edu/nll/?p=34251

Lindley, Mark. (2001). Well-tempered clavier. *The New Grove Dictionary of Music and Musicians*, second edition, Stanley Sadie and John Tyrrell (eds). London: Macmillan Publishers.

Liu J.; Wang N.; Li J.; Shi B. & Wang H. (2009). Frequency distribution of synchronized spontaneous otoacoustic emissions showing sex-dependent differences and asymmetry between ears in 2- to 4- day-old neonates. *International Journal of Pediatric Otorhino-laryngology,* 73(5), 731-736.

Maghbouleh, A. (1998). Tobi accent type recognition. In *Proceedings of ICSLP* Paper 0632.

Martin, P. (1981) Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne, *12e Journées d'Etude sur la Parole*, SFA, Montréal, 1981.

Mertens, P. (2004). The Prosogram: semi-automatic transcription of prosody based on a tonal perception model. In *Proceedings of the 2nd International Conference on Speech Prosody* Nara (Japan) pp 549-552.

Mertens, P. (2018). Prosogram, v 2.15. Pitch contour stylization based on a tonal perception model [webpage: https://sites.google.com/site/prosogram/home]

Mertens P. & d'Alessandro C. (1995). Pitch contour stylization using a tonal perception model. *Proceedings of the 13th International Congress of Phonetic Sciences* vol. 4, pp. 228-231.

Mixdorff, H. -J. (1999). A novel approach to the fully automated extraction of Fujisaki model parameters. In *Proceedings of ICASSP* 1999. pp. 1281-1284.

Moore, B. C. J. & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America* 74, 750-753.

Moore, B. C. J. & Glasberg, B. R. (1996). A revision of Zwicker's loudness model. *Acta Acustica,* 82, 335-345.

Morel, A. (1980). *Codage des sons dans le corps genouille médian du chat: évaluation de l'organisation tonotopique de ses différents noyaux*. PhD dissertation. Université de Lausanne, Juris, Zurich.

Morest, D. K. (1965). The laminar structure of the medial geniculate body of the cat. *Journal of Anatomy* 99, 143-160.

Nolan, F. (2003). Intonational equivalence: an experimental evaluation of pitch scales. in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, pp. 771-774.

Nooteboom, S. (1999). The prosody of speech melody and rhythm. In Hardcastle, W. J. & Laver, J., eds., *The Handbook of Phonetic Sciences*. London: Blackwell. pp. 640-673

O'Shaughnessy, D. (1987). *Speech Communication: Human and Machine*. Reading, Mass.: Addison-Wesley: 150. [ISBN 978-0-201-16520-3]

Paeschke, A., & Sendlmeier, W. F. (2000). Prosodic characteristics of emotional speech: measurements of fundamental frequency movements. In *Proceedings of the ISCA-Workshop On Speech and Emotion.* Belfast, Ireland, pp. 75-80.

Rossi M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole." *Phonetica,* 23, 1-33.

Silverman, K. (1986). Fo segmental cues depend on intonation: The case of the rise after voiced stops. *Phonetica,* 43(1-3), 76-91.

Steele, J. (1779). *Prosodia Rationalis: or, an Essay towards Establishing the Melody and Measure of Speech, to be Expressed and Perpetuated by Peculiar Symbols,* 2nd edn. London: J. Nichols.

Stevens, S., Volkman, J., & Newman, E. (1937). A scale for the measurement of the psychological magnitude of pitch. *Journal of the Acoustical Society of America,* 8, 185–190.

Taylor, P. (1995). The rise/fall/connection model of intonation. *Speech Communication*, 15(1-2), 169–186.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America,* 88, 97-100.

Traunmüller, H. (1997). Auditory scales of frequency representation. [webpage] http://www2.ling.su.se/staff/hartmut/bark.htm

Umesh, S., Cohen, L. & Nelson, D. (1999). Fitting the mel-scale. In *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, vol. 1, Phoenix, Arizona, USA, March 1999, pp. 217–220.

Véronis, J., Hirst, D. J. & Ide, N. (1994). NL and speech in the MULTEXT project. In *Proceedings of AAAI Workshop on Integration of Natural Language and Speech*, 72–78, Seattle, USA.

Wightman, C. & Campbell, N. (1995). Improved labeling of prosodic structure. In *IEEE Transactions on Speech and Audio Processing*.

Warren, P. & Calhoun, S. (2019). Intonation. *This volume.* Chapter 8.

Wikipedia (2018) Pitch detection algorithm [web page: https://en.wikipedia.org/wiki/Pitch_detection_algorithm, accessed 2018-09]

Wright, A. A., Rivera, J. J., Hulse, S. H., Shyan, M. & Neiworth, J. J. (2000). Music perception and octave generalization in rhesus monkeys. *Journal of Experimental Psychology Gen* 129 (3), pp. 291- 307

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenz-gruppen). *Journal of the Acoustical Society of America,* 33, p. 248.

Zwirner, E. & Zwirner, Z.K. (1937) Über das Hören und Messen des Sprachmelodie, *Achiv für vergleichende Phonetik* 1, pp. 35-47.