(To appear in: Robert Podesva and Devyani Sharma (eds.) *Research methods in linguistics*. Cambridge University Press)

Chapter 17: Acoustic analysis

Paul Boersma, 2 February 2013

1. Introduction

Acoustic analysis, once a method used primarily within the domain of phonetics, has become an increasingly necessary skill across the field of linguistics. To name just a few examples: phonologists sometimes appeal to acoustic data to substantiate theoretical arguments, sociolinguists tend to characterize vowel shifts and mergers in terms of their acoustic properties, and psycholinguists frequently draw on acoustic analysis techniques to construct stimuli for experiments.

The analysis of acoustic signals is mainly performed with the help of generally available software. Because of its capability of creating publication-quality graphics, the pictures in this chapter were made with Praat (Boersma & Weenink 1992–2013), a general set of tools for analysing, synthesizing and manipulating speech and other sounds bundled into a single integrated computer program. Praat is available free of charge for all current major computer platforms (nowadays MacOS, Windows, Linux) and is continually updated to accommodate new operating system developments and new analysis methods.

Graphical software allows us to perform acoustic analysis by inspecting visualized speech. The types of visualization addressed in the present chapter are the waveform, the pitch curve, the intensity curve, the spectrum, the spectrogram, and formant tracks. These types of visualization will be seen to help in measuring the following articulatory, acoustic, and auditory quantities: glottal period, resonance frequencies, pitch, duration, intensity, noisiness, and place of articulation. Examples of practical uses for each of these measures appear throughout this chapter.

2. The waveform

The *waveform* of a sound is the direct visualization of the sound as recorded by a microphone and represents the air pressure as a function of time. In the waveform one can directly see when there is silence and how long the utterances are, but one can also infer from the waveform many acoustic properties of speech, among which periodicity, intensity and spectral qualities.

2.1 The waveform of a vowel

Figure 1 shows a part of the waveform of a recording of a token of the Dutch vowel /i/, as spoken by the present author in 1997. The horizontal axis represents the time as expressed in the number of seconds that have elapsed since the start of the recording. The vertical scale represents the air pressure recorded by the microphone. The vocal folds close approximately at the times g_1, g_2, g_3, g_4, and g_5. These are the times at which the folds hit one another, causing a loud clapping noise at the glottis, which leads to resonances in the vocal tract. Thus, the clap at g_2 causes strongly rising and falling air pressures just after g_2, and these resonances gradually die out, leading to smaller rising and falling air pressures as the time proceeds towards g_3. Just before g_3, the air pressure becomes strongly positive, which corresponds to the air being compressed in the glottis just before the vocal folds touch each other again.

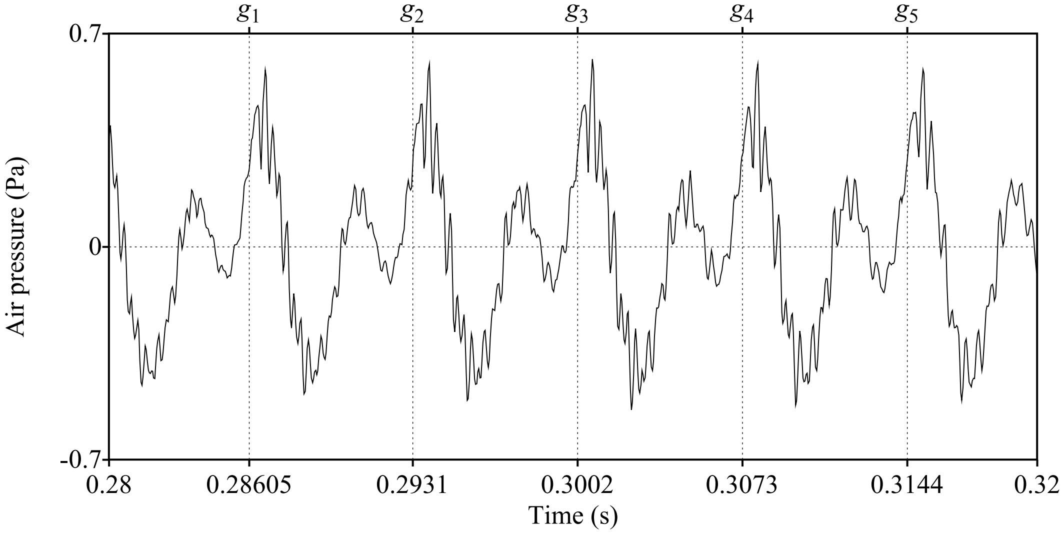

Fig. 1. Waveform of several periods of the Dutch vowel /i/, illustrating glottal fold vibration.

The distance in time between consecutive vocal fold closures is the *vocal fold vibration period*, or T0; estimates of T0 are $g_2 - g_1$ = 0.2931 – 0.28605 = 0.00705 s and $g_3 - g_2$ = 0.3002 – 0.2931 = 0.0071 s. Therefore, approximately 1/0.0071 = 141 of these periods fit in one second, so that the *vocal fold vibration frequency* must be about 141 Hz (hertz). This frequency is an important quantity in speech research, because humans tend to be able to hear it and languages therefore employ it in implementing tone and intonation. More specifically, the human auditory system has a periodicity detector, or *pitch detector*, which recognizes recurring waveshapes: for the sound in the figure, humans tend to perceive a *pitch* of 141 Hz.

While pitch is the frequency with which the whole wave shape repeats itself, the waveform also contains other frequencies, namely those associated with the sine-like waves that represent resonances of the vocal tract. Figure 2 zooms in on Fig. 1 and shows the period of a slow resonance: the duration of one vibration of the slow sine wave can be measured as the time between two consecutive peaks, e.g. $s_2 - s_1$ = 0.2977 – 0.2939 = 0.0038 s. It can also be measured as the time between two consecutive valleys, e.g. $s_4 - s_3$ = 0.3063 – 0.3029 = 0.0034 s. Therefore, the period of the slow resonance is approximately 0.0036 seconds long. Approximately 1/0.0036 = 278 of these periods fit in one second, so that the *slow resonance frequency* must be estimated as 278 Hz. Such values are typical for the Dutch high vowels /i/, /y/ and /u/. The slow resonance frequency is another important quantity in speech research, because humans can hear it and languages therefore employ it in implementing the phonological feature of vowel height. More specifically, the human auditory system has a *spectral analysis system* (namely the basilar membrane in the inner ear and the neural circuitry emanating from it), which dissects the incoming sound into its component sine waves: for the sound in the figure, humans tend to perceive a *first formant* of 278 Hz.

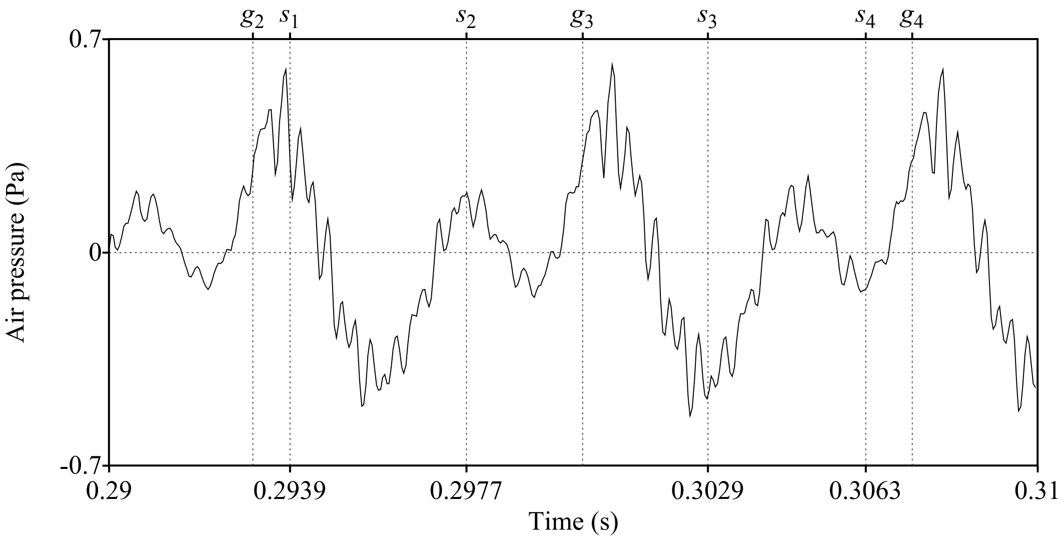

Fig. 2. Waveform of several periods of the Dutch vowel /i/, illustrating the first formant.

The third phenomenon that the waveform shows is a *rapid resonance frequency*. Figure 3 zooms in a bit more than Fig. 2, and a rapidly vibrating resonance becomes clearly visible. Six periods of it lie between r_1 and r_7, which are 0.30270 - 0.30084 = 0.00184 seconds apart. Each period therefore lasts 0.00184 / 6 = 0.000307 s, so that 1 / 0.000307 = 3200 of these periods fit in one second. The rapid resonance frequency, which humans tend to perceive as the *second formant*, is therefore 3200 Hz. This value is typical for the Dutch vowel /i/ (acoustically, this resonance corresponds to the third and fourth formants of the vocal tract; the true second formant may lie around 2200 Hz but is weak and not visible in the figure).

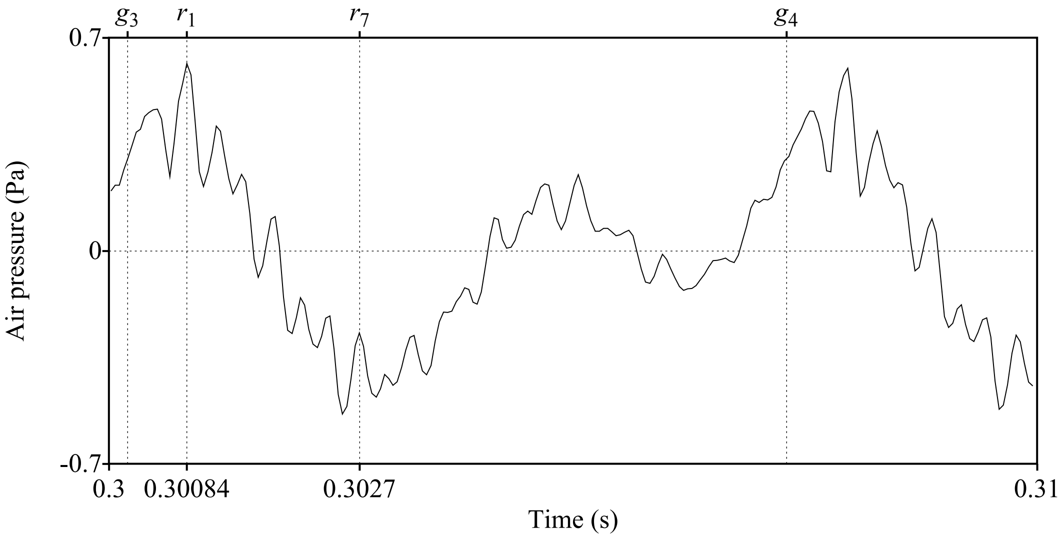

Fig. 3. Waveform of several periods of the Dutch vowel /i/, illustrating the second formant.

The choice for the vowel /i/ as the example for the present section was informed by the large distance between the first and second formant. In other vowels than /i/, however, this distance tends to be much smaller. Figure 4 shows a part of the waveform of a token of Dutch /a/, with time markers every 0.001 seconds. Some consecutive major positive peaks and some consecutive major negative peaks are just over 0.001 seconds apart, and some consecutive peaks or valleys are a bit less than 0.001 seconds apart, but we cannot see any well-defined sine waves as we could with /i/. The waveform is made up of a slower and a faster resonance whose periods are a bit above and a bit below 0.001 seconds, respectively, but their periods are so close together that their sine waves visually interfere, so that the two resonances are hard to distinguish from each other visually. For reasons such as this, vowel formants are usually investigated not with the help of waveforms but with the help of spectral techniques, as described in sections 5 and 6.

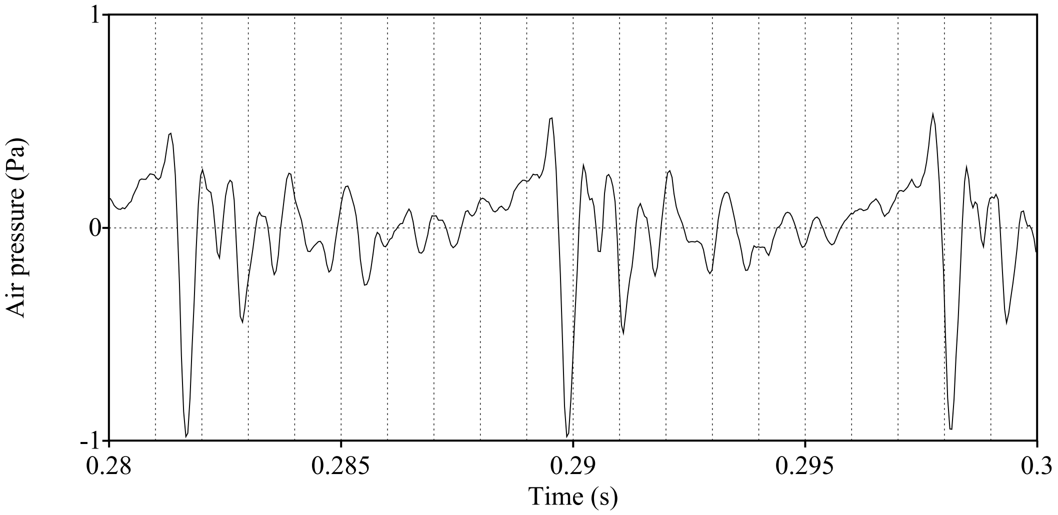

Fig. 4. Waveform of several periods of the Dutch vowel /a/, illustrating mangled formants.

A fourth type of property visible in the waveform is *duration*. Figure 5 shows the whole /i/ of Figs. 1 through 3. The vowel starts at a time of d_1 = 0.015 seconds and ends at a time of d_2 = 0.911 seconds, from which we can conclude that its duration is 0.911 – 0.015 = 0.896 seconds. Duration is an important quantity in speech research, because humans have a mechanism for measuring the duration over which a signal stays approximately stationary in terms of other percepts (such as the first and second formant here) and languages therefore employ duration as the major cue to the phonological length of vowels and consonants; moreover, duration is a cue to many other phonological elements (including stress, obstruent voicing, and vowel height) and to paralinguistic features of speech.

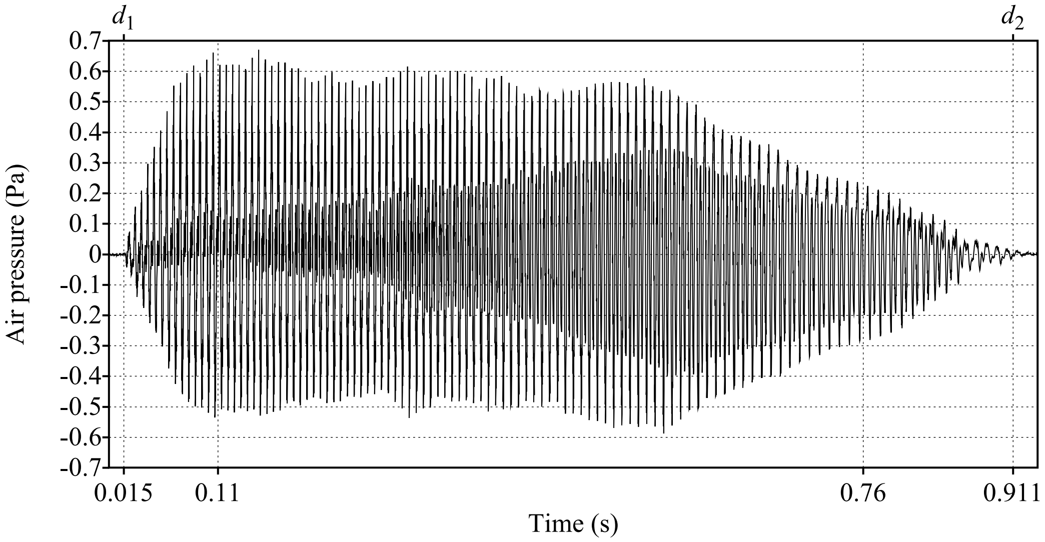

Fig. 5. Waveform of a whole Dutch /i/, illustrating duration and intensity.

The fifth acoustic quantity visible in the waveform is *intensity*. In Fig. 5, the top-to-top amplitude of the sound (peak minus valley) at a time of 0.11 seconds is more than 1.1 Pa, whereas at a time of 0.76 seconds it has fallen to approximately 0.5 Pa. Now, the *absolute amplitudes* of this sound at the time and place of recording are probably different from 1.1 and 0.5 Pa, because the gain of the recording was not calibrated (the fake numbers in Pa in the figure were computed from the sound file, where the minimum and maximum representable values were arbitrarily assigned the values of –1 and +1 Pa). However, the *relative amplitude* of the different parts of the sound (i.e. a fall by a factor of 2.2 between 0.11 and 0.76 seconds) is reliable, assuming that the speaker kept a constant distance to the microphone and nobody turned the gain control during the recording. Relative intensity is an important quantity in speech research, because it contributes to the perception of phonological phenomena such as stress, stridency, manner, voicing, and nasality.

2.2 Other waveforms

Figure 6 shows the waveform of the voiceless palatal plosive in [aca]. Between the two vowels, the figure indicates a silence with a duration of 140 ms, followed by a release burst with a duration of 24 ms. The voicing of the vowel starts right after the burst ends. This is a non-aspirated plosive, but the *voice onset time*, which is defined as the time at the start of voicing minus the time at the release, is rather positive, namely 24 milliseconds. Of all the acoustic measurements discussed in this chapter, voice onset time is one of the few that can best be measured from the waveform.

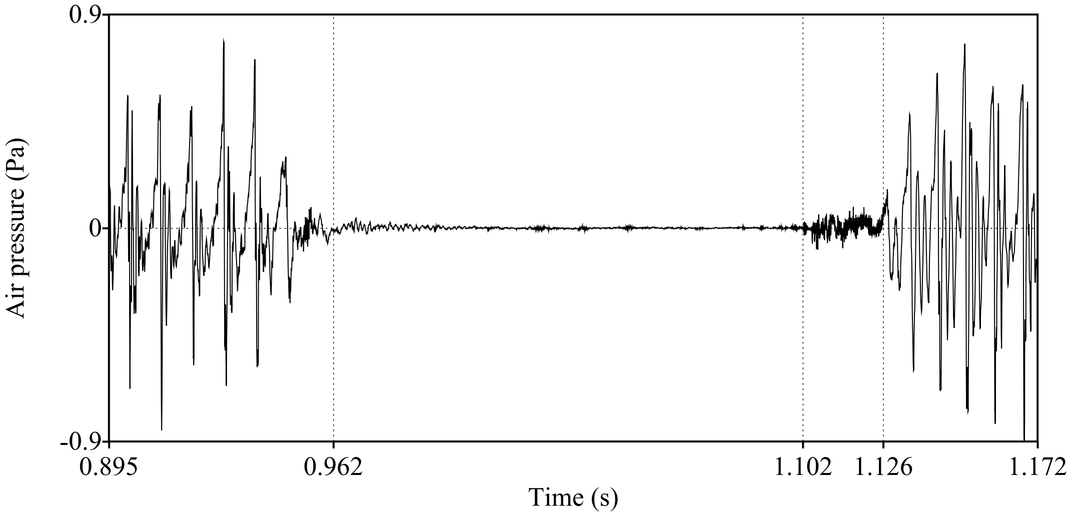

Fig. 6. Waveform of the voiceless palatal plosive in [aca], illustrating silence and release burst.

Figure 7 shows the waveform of the voiceless palatal fricative in [aça]. The fricative noise lasts 211 milliseconds and can be seen from the large number of times the waveform crosses the 0 Pa line every millisecond.

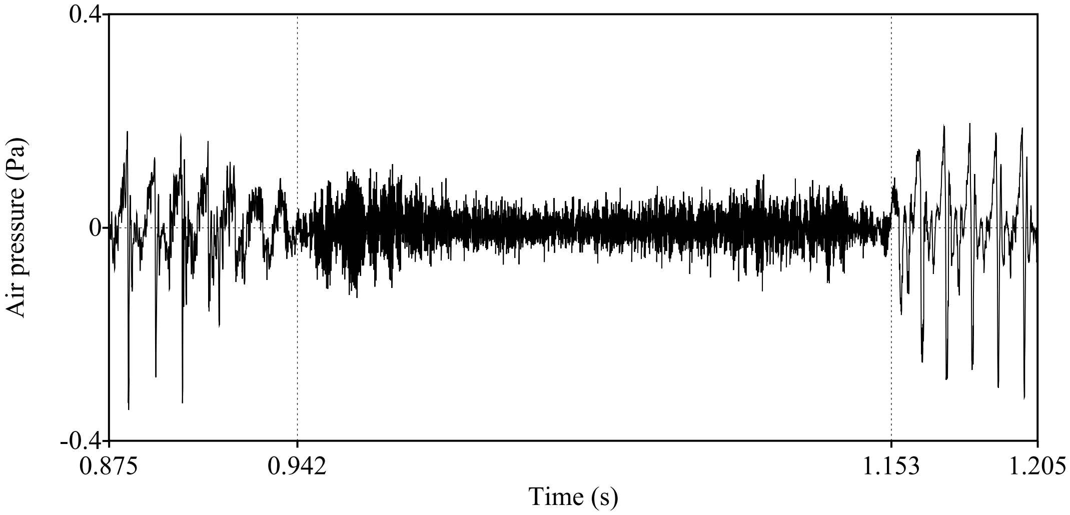

Fig. 7. Waveform of the voiceless palatal fricative in [aça], illustrating the many zero crossings.

Finally, Fig. 8 shows the waveform of the alveolar trill in [ara]. While the vocal folds continue to vibrate during the trill, four tongue tip closures cause the amplitude of the waveform to fall towards zero at the indicated four points in time. Three periods of vibration fit into 1.065 – 0.947 = 0.118 seconds, so that there are approximately 3 / 0.118 = 25 tongue-tip vibrations per second.

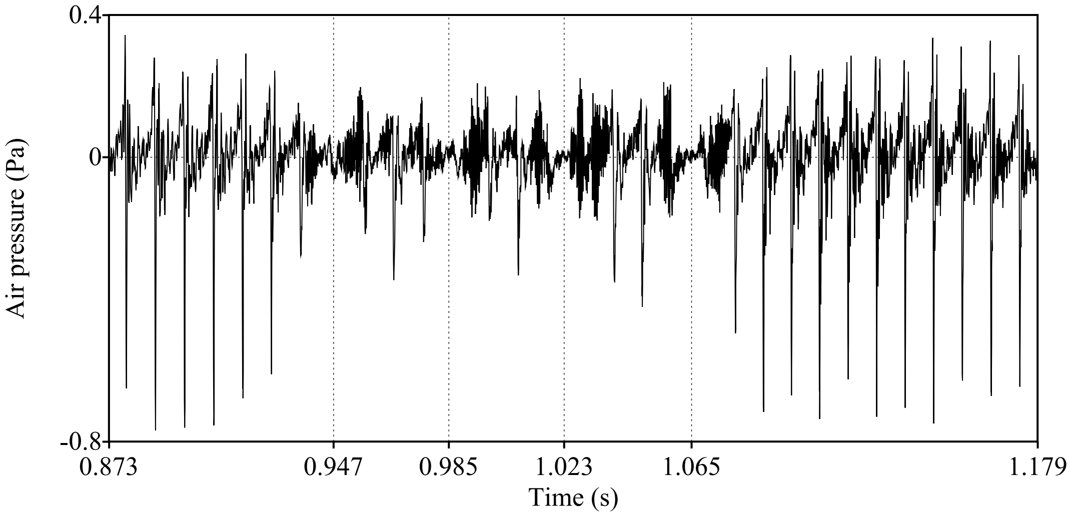

Fig. 8. Waveform of the alveolar trill in [ara], illustrating four passive tongue-tip closures.

2.3 Applications and limitations of waveform inspection

The usefulness of the waveform for acoustic research is that it is basic, shows whether there is speech or silence, and constitutes the main source of information on voice onset time. Most other acoustic phenomena are slightly to strongly easier to investigate with different types of visualization, such as pitch curves, spectra, and spectrograms, all of which are discussed in the following sections.

3. Periodicity analysis

In Fig. 1 the glottal fold vibration frequency was determined by inspecting and measuring the waveform. If you want to determine the tonal pattern or intonation contour of a whole utterance, such a procedure is impractical. Fortunately, phonetic analysis software provides automated pitch measurement techniques.

3.1 Automated pitch measurement techniques

Most automated pitch measurement techniques are based on the self-similarity of the waveform. In Fig. 1, for instance, the 7.1-ms long part from 0.2931 to 0.3002 seconds is extremely similar to the adjacent 7.1-ms long part from 0.3002 to 0.3073 seconds, whereas, for instance, the 5.0-ms long part from 0.2952 to 0.3002 is quite dissimilar from its adjacent 5.0-ms long part from 0.3002 to 0.3052 seconds. Figure 9, which copies these and several other parts of Fig. 1 just before and just after 0.3002 seconds, illustrates these similarity verdicts. If we want to make a guess about the “true” glottal period at 0.3002 seconds, then 7.1 ms seems to be a much better candidate than 5.0 ms. In fact, 7.1 ms looks a much better candidate than 6.5 or 8.0 or 10.0 ms as well (in Fig. 9), and if one does the computations, then 7.1 ms turns out to be an even slightly better candidate than 7.0 or 7.2 ms. An automated pitch measurement technique based on *cross-correlation* (Talkin 1995) will now say that at 0.3002 seconds the glottal period is 7.1 ms and that therefore the pitch is 1/0.0071 = 141 Hz. Analogous statements can be made about every time point in Fig. 1: for every time point it is possible to look backward and forward in time and to determine how similar the immediate future is to the immediate past.

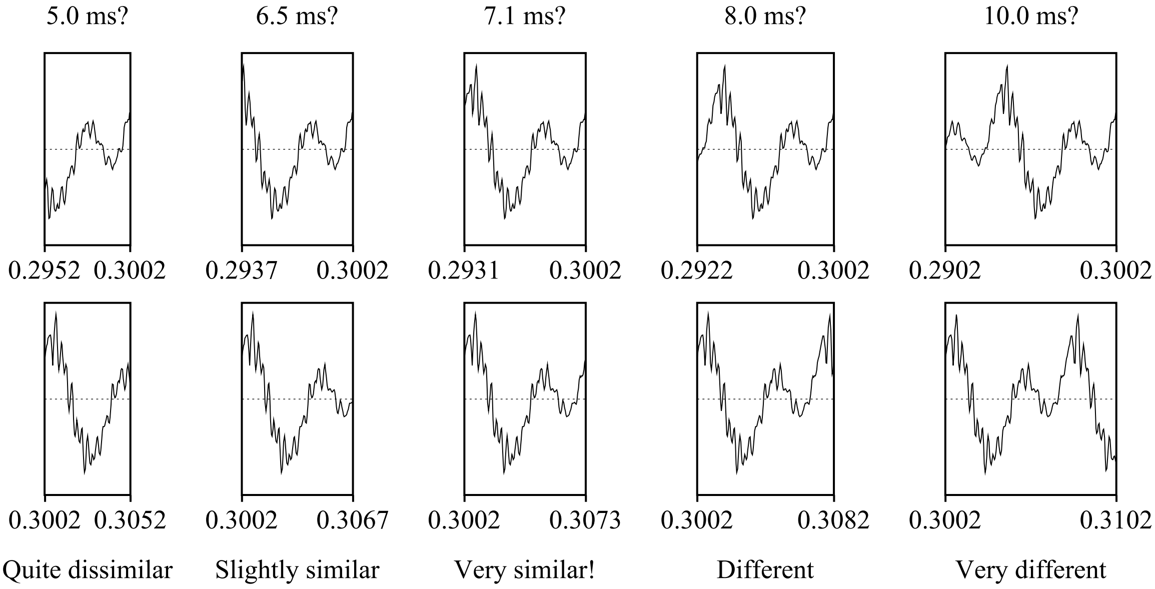

Fig. 9. Determining the pitch of the sound in Fig. 1 at a time of 0.3002 seconds (cross-correlation method). The top row shows parts of the sound just before that time, and the bottom row shows equally long parts just after. The two parts look most similar if they are 7.1 ms long.

A general property of automated acoustic measurements is the use of an *analysis window*. In the example just mentioned, the F0 at 0.3002 seconds cannot be determined by looking at what happens at that time point, but it can only be determined by looking some time into the past and some time into the future. If you want to detect any F0 between 100 Hz and 500 Hz, you will have to look into the past and future for at least 2.0 ms and at most 10.0 ms; so for the lowest F0 you want to detect (100 Hz), you will have to consider an analysis window of 20.0 ms (i.e. 10.0 ms in both directions).

A related technique is *autocorrelation*. In the example above, while cross-correlation works with windows whose lengths vary between 4.0 and 20.0 ms, the autocorrelation method works by looking at a time window with a constant length of 30.0 ms and computing the similarities of all amplitudes spaced apart within that window by times between 2.0 and 10.0 ms. Autocorrelation methods have a bad reputation in the literature, because early versions could produce much too high F0 estimates (typically, one octave higher than the true F0); however, this problem was solved by Boersma's (1993) *unbiased* autocorrelation method.

3.2 What automated pitch measurements look like

In the [i] vowel of Fig. 5 one can see that the glottal periods get longer towards the end of the vowel. Apparently, the glottal fold vibration frequency falls during the course of this vowel. The autocorrelation method tracks the development of this frequency in the way shown in Fig. 10. The figure does not show pitch values from 0.8 seconds on; apparently, the automated pitch measurement method considers the signal insufficiently voiced or too quiet in that region.

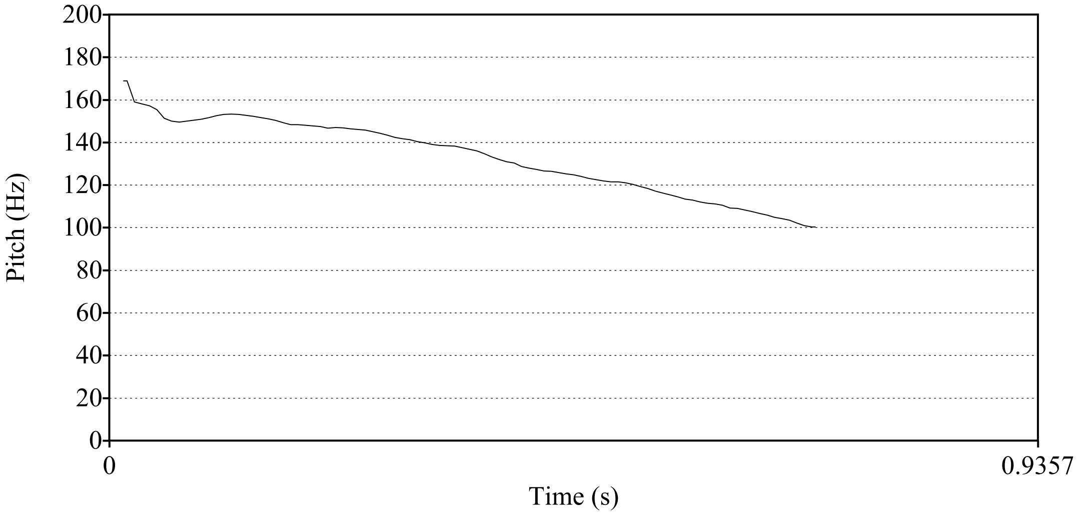

Fig. 10. Pitch curve for the [i] vowel of Fig. 5.

Loosely, pitch measurement techniques can be said to measure either an articulatory phenomenon, i.e. the glottal fold vibration frequency, or a mathematical phenomenon, i.e. (near-)periodicity, or an auditory phenomenon, i.e. the perceived pitch. What the autocorrelation method of Fig. 10 measures is closest to the perceived pitch (without the niceties of experimental psychoacoustic results with nonperiodic signals to which humans can nevertheless assign pitch), which is appropriate for intonation or tone research (Boersma 1993). The cross-correlation method has the disadvantage of making intonation or tone mistakes in the presence of noise, but comes closer to measuring the actual glottal pulses, and is therefore appropriate for measuring aspects of voice quality, such as jitter, shimmer and harmonics-to-noise ratio (Boersma 1993, 2009).

3.3 Limitations of automated pitch measurements

In automated pitch measurements, several things can go wrong. In Fig. 9, 7.1 ms is clearly the best candidate for the glottal period, but from Figs. 2 and 3 we can see that the same sound contains sine waves that have peaks every 3.8 or 0.307 ms, so that 3.8 ms (the period of F1) and 0.307 ms (the period of F2) are pitch candidates that at least fall in the "fairly similar" rubric. These extraneous pitch candidates are normally overruled by the much better matching true glottal period of 7.1 ms, but in voiceless parts of the sound, where there is no true glottal period, these formants might become the best pitch candidates, especially if there is echo in the background. As a result, a pitch measurement procedure might hallucinate that these formants are pitches, and in practice we see that pitch analysis tools will indeed show spurious pitches in voiceless stretches. To ameliorate this problem, pitch analysis tools often allow you to set a maximum pitch value above which the tool will ignore any pitch candidates. This "pitch ceiling" was 300 Hz for the male voice in Fig. 10, and you can set it to 500 Hz for female voices. This works reasonably unless people are yelling, singing or otherwise stretching their voice.

The situation is even worse at the lower side of the pitch range. In Fig. 1 we can see that 142 ms is an almost equally viable candidate for a period as 71 ms is, i.e. the part from 0.2860 to 0.3002 seconds is very similar to the part from 0.3002 to 0.3144 seconds. This means that 1 / 142 ms = 70.5 Hz is a virtually equally good pitch candidate as the true pitch of 141 Hz is. In order to prevent the pitch analysis tool from making an "octave error" (i.e. proposing a pitch of 70.5 Hz instead of 141 Hz), the pitch analysis tool has to have a controlled small bias in favour of higher frequencies. In noisy situations, however, this bias might not suffice, i.e. the similarity over 142 ms might be greater than the similarity over 71 ms just by chance. To ameliorate this problem, pitch analysis tools often allow you to specify a minimum pitch value below which the tool will not look for pitch candidates. This “pitch floor” was 75 Hz for the male voice in Fig. 10, and you can set it to 100 Hz for female voices. For creaky voice you should set the pitch floor much lower than 75 Hz, e.g. to 40 Hz.

The importance of a sufficiently high pitch floor in the presence of noise has been confirmed experimentally by Deliyski et al. (2005), who investigated the quality of several pitch measurement methods in 10 situations: 2 speaker sexes (male, female) times 5 levels of background noise (fan, 60 Hz, white, talk, traffic). In 9 of these 12 situations, including all male conditions, the autocorrelation algorithm of Fig. 10 outperformed another algorithm, but in the other 3 situations it performed worse, making many mistakes; all three situations were with female voices, and what probably contributed to the mistakes was the fact that the authors had set the pitch floor as low as 70 Hz. Speech researchers are therefore advised to take this setting seriously.

4. Intensity analysis

In the discussion of Fig. 5 it was claimed that the course of the intensity could be seen from the waveform. However, this is just an approximation. If we take the height of the peaks in Fig. 5 as a criterion, we have to conclude that the highest intensity lies around 0.15 seconds. But “intensity” refers to the period-averaged power in the signal, so it is also important to look at what happens between the peaks. Indeed we see in Fig. 5 that around 0.6 seconds the signal is “thicker”, i.e. the peaks may be lower but the amplitudes between the peaks are greater than at 0.15 seconds. In order to measure exactly the course of the power in the signal, automated intensity measurements can help.

4.1 An automated intensity measurement technique

Figure 11 shows how the intensity curve of the sound in Fig. 5 can be determined. We like the end result to be expressed at every moment in time as a number of dB above the human auditory intensity threshold. As this threshold is defined as a pressure of 0.00002 Pa, we first divide the pressure curve of Fig. 5 by this value of 0.00002 Pa. The result is in Fig. 11a; the curve is identical to Fig. 5 except for the vertical scaling. The next step comes from realizing that the power in a pressure signal is proportionate to the square of the pressure (physically, one can say that power equals air pressure times volume velocity, and if you increase the signal strength in such a way that the air pressure increases by a factor of ten, the air particles will also speed up by a factor of ten). Figure 11b therefore shows the square of Fig. 11a; this cannot yet be called the intensity of the signal, because there are still within-period fluctuations. To smooth these away, we *convolve* the signal with the unit-area Gaussian kernel of Fig. 11c, yielding the smoothed intensity curve of Fig. 11d. The height of this curve is less than that of the peaks in Fig. 11b, because in Fig. 11d the original peaks have been averaged out with their surrounding valleys; the *area* under the curve in Fig. 11d, however, is still the same as the area under the curve in Fig. 11b. Finally, it is usual to draw the intensity curve along a logarithmic vertical axis, as in Fig. 11e, where every factor of ten in intensity is awarded an equal part of the vertical space. These intensities, which are still taken relative to the auditory threshold (if the signal is calibrated), can be straightforwardly translated to values in dB, with e.g. a threshold-relative intensity of 10^8 corresponding to an intensity level of 80 dB, as illustrated in Fig. 11e.

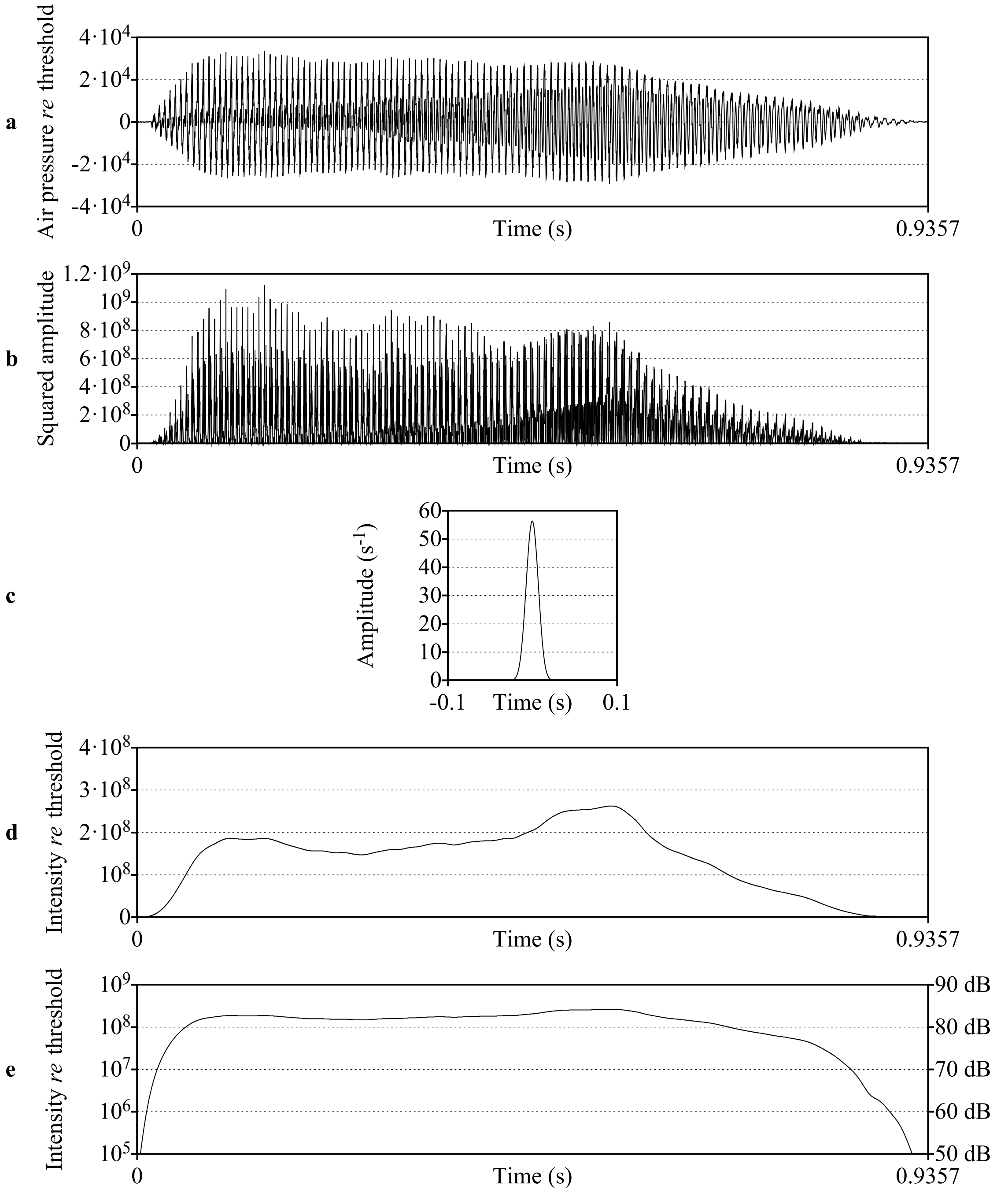

Fig. 11. Determining the intensity curve for the [i] vowel of Fig. 5: (a) the original sound, as measured relative to the auditory threshold; (b) the square of this; (c) the Gaussian smoothing kernel, on the same time scale as the sound; (d) the intensity curve, computed as the convolution of the squared amplitude and the Gaussian; and (e) the intensity curve along a logarithmic scale.

In the curves of Figs. 11d and 11e, we see that the intensity peak around 0.6 seconds is indeed stronger than that around 0.15 seconds. Apparently, the “thickness” in the waveform of Fig. 11a around 0.6 seconds outweighs the height of the peaks in the waveform around 0.15

seconds. Thus, automated intensity measurement techniques can provide precision that the human eye cannot.

5. Spectral analysis

The dissection of a sound into its component sine waves, which I illustrated with the waveforms of Figs. 2 and 4, can be automated.

5.1 An automated spectral analysis technique

Figure 12 shows how the periodic [i]-like speech sound of Fig. 2 can be approximated as the sum of six sine waves.

The top picture shows exactly two periods of this voiced sound. The sound marked “1+2” is a rough approximation of these two periods: one can see that it follows all slow movements of the original sound. This sound “1+2” is composed of two sine waves, namely the sound marked “1” and the sound marked “2” in the figure. The sound marked “1” is a sine wave with the same frequency as the glottal fold vibration, i.e., it has a frequency of F0 = 141.346 Hz. The sound marked “2” is a sine wave with exactly twice that frequency, i.e., it is the *second harmonic* of F0 and has a frequency of 2F0 = 282.692 Hz. In the figure it can be seen that this second harmonic has an amplitude slightly greater than that of the first harmonic (the sound marked “1”). When we add these two sounds to each other, we obtain the sound “1+2”; for instance, sound “1” starts with a negative value (at 0.297020 seconds) whereas sound “2” starts with a positive value, and in the sound “1+2” these negative and positive values add up to approximately zero, which is the value of “1+2” at its start. The summed curve “1+2” is computed in this way from the curves “1” and “2” at every time point.

When we include higher frequencies, the match between the summed sine waves and the original sound improves. When the 15th, 22nd, 23rd and 24th harmonics, with amplitudes as shown in the figure are added to the sound “1+2”, we obtain the sound that is marked as “1+2+15+22+23+24” in the figure. This summed sound is very close to the original, both in its wave shape and in how it sounds to the human ear. Apparently, these two periods of the [i]-like sound of Fig. 2 can be well approximated as the sum of six sine waves.

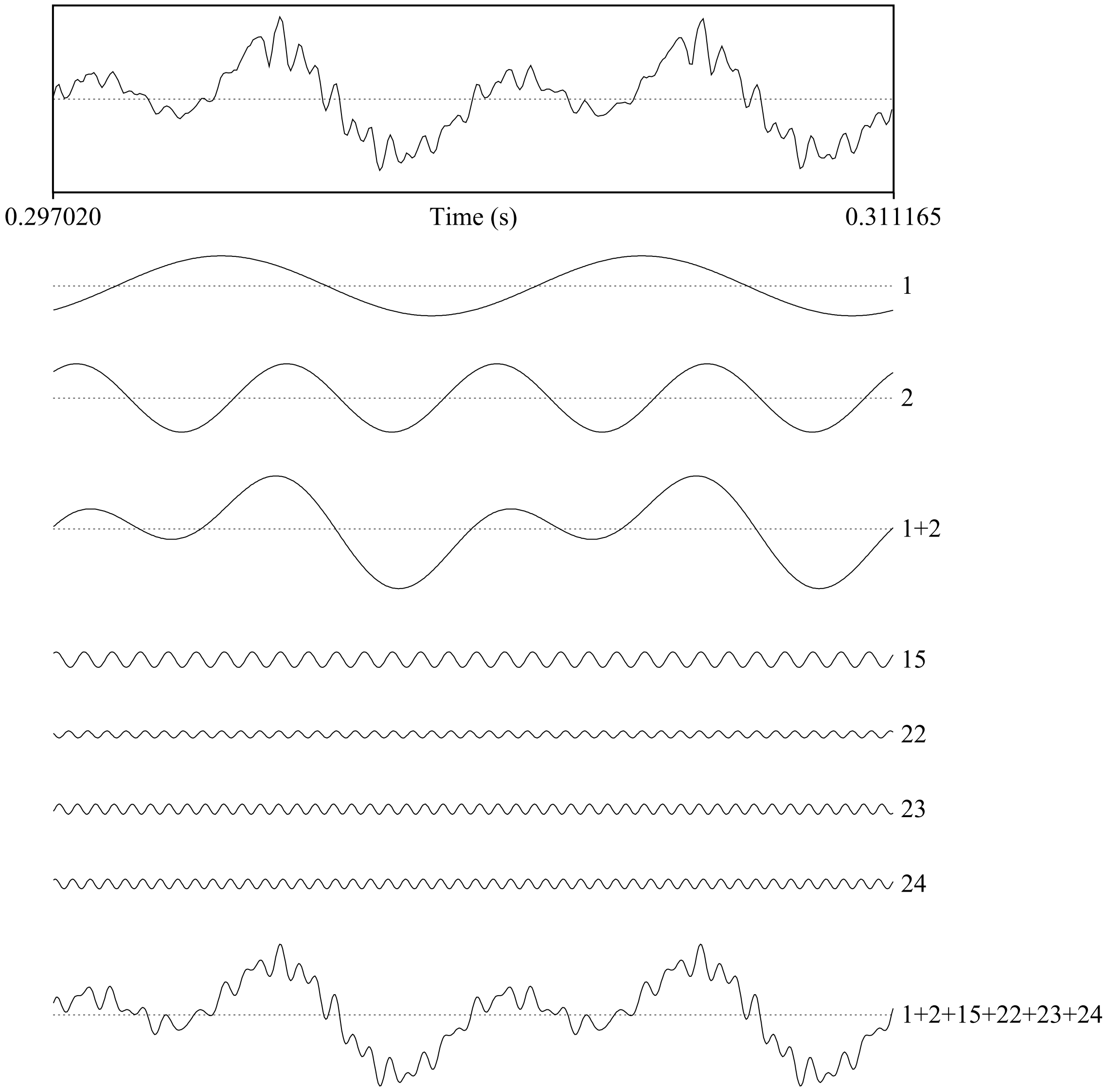

Fig. 12. Splitting up two periods of the [i] vowel of Fig. 5 into six harmonics. At the top the original sound. The rough features of the original sound are reconstructed by adding the first harmonic (1) and the second harmonic (2) to each other (1+2). When we add the 15th, 22nd, 23rd and 24th harmonics to this, the original waveshape is approximated even more closely (bottom).

This method of approximating a periodic sound as a sum of sine waves was developed by Fourier (1822). I used Fourier's formulas to determine the amplitudes of the six waves in Fig. 12, as well as to determine the *phase*, i.e. the horizontal time shift, of each sine wave in Fig. 12. When this is done for every harmonic, not just for the six harmonics in Fig. 12, we obtain Fig. 13, which shows the strength of each frequency component up to 5000 Hz. In spectral pictures like these, the horizontal axis represents frequency rather than time. We see that the 1st, 2nd, 15th, 22nd, 23rd, and 24th harmonics (marked along the top of the picture) are strong, but that most other harmonics also play a role; apparently, the match in Fig.12 would have been even better if we had added harmonics 3, 14, 21, 25 and so on. In between the harmonics, Fig. 13 shows that the sound contains components of nonzero amplitude (a zero amplitude would have shown up as $-\infty$ dB in the figure); this indicates that the sound is not perfectly periodic, as can be confirmed in the waveform of Fig. 12 (top).

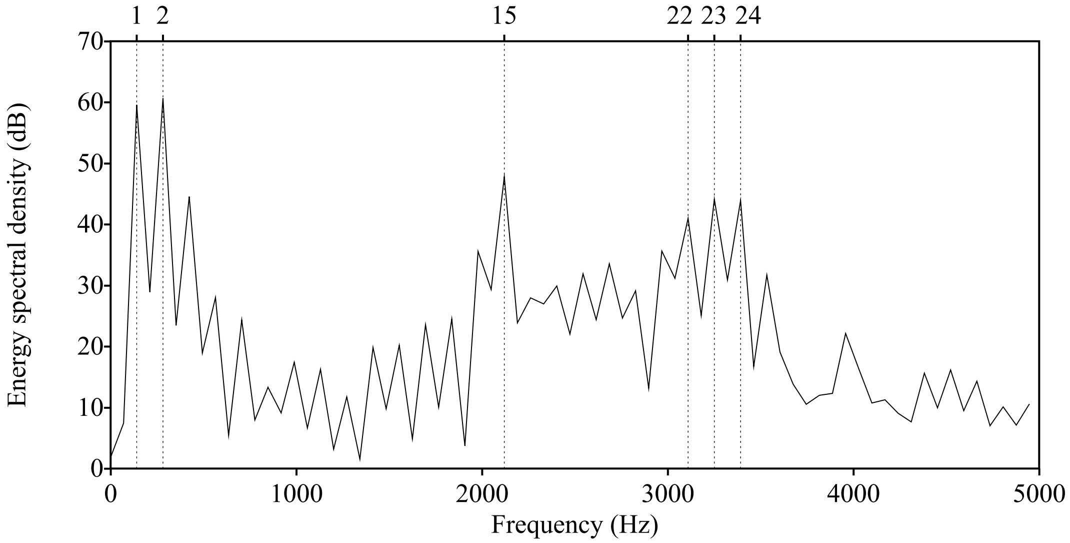

Fig. 13. The Fourier spectrum of the two-period [i]-like sound of Fig. 12.

5.2 What automated spectral measurements look like

In the case of Fig. 13 we had a longer sound from which we took exactly two periods, leading to a spectral shape with peaks at every harmonic and valleys exactly in between the peaks. More usually, you will select a stationary part of a speech sound and ask your acoustic analysis program to provide an average spectrum over that selection.

The [i]-like sound of Fig. 5 is stationary with respect to its formants, i.e., its formants do not change auditorily throughout the vowel; its F0, on the other hand, steadily falls. If we are interested in the formants rather than in the pitch, we can do a spectral analysis on the whole sound. Figure 14 shows the spectrum of this longer sound. Pronouncing the sound with a moving pitch wipes out most of the harmonic structure of Fig. 13 and allows us to see the formants quite well in the spectrum: Fig. 14 shows a first formant at 300 Hz with a strength of 60 dB, an F2 at 2150 Hz with a strength 15 dB lower than that, and an F3 and F4 at 3000 and 3400 Hz, respectively (these last two were not easy to distinguish in Fig. 13).

Many more spectra of speech sounds, with explanations of how they come about articulatorily, can be found in Stevens (1998).

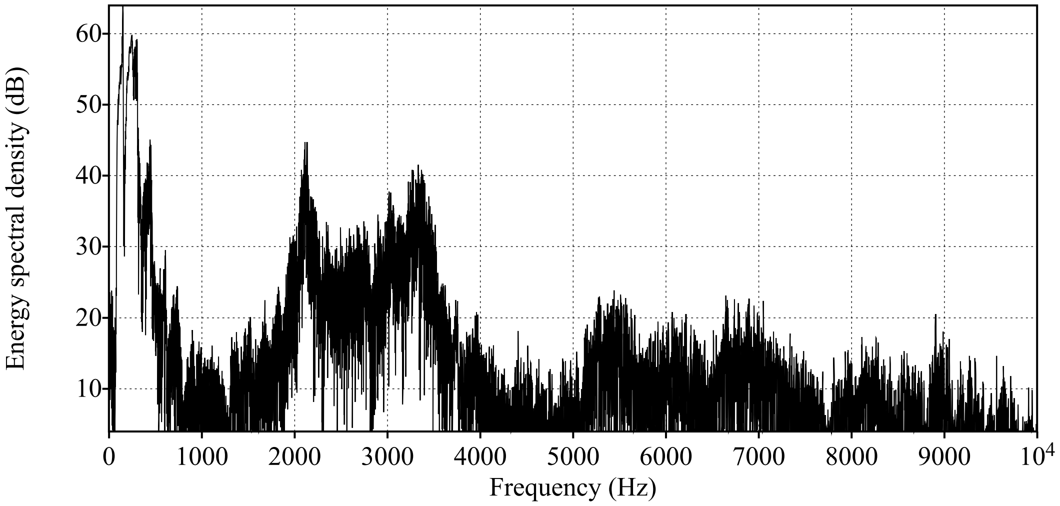

Fig. 14. Spectrum of the vowel [i].

5.3 Applications and limitations of automated spectral measurements

Fourier's method of measuring the spectral content of speech sounds is especially appropriate for sounds with a stationary part, such as fricatives, vowels (monophthongs) and nasal consonants, but is less appropriate for speech sounds whose characteristics involve crucial dynamic changes, such as plosives, diphthongs, and trills. For those, the technique has to be extended to include change in time, as is done in the spectrogram.

6. The spectrogram

The spectrogram is the workhorse of speech visualization. It is employed with equal enthusiasm in textbooks (e.g. Ladefoged & Disner 2012) and handbooks (e.g. Maddieson & Ladefoged 1996). The spectrogram shows the frequency contents of a sound as a function of time, and thereby follows the capabilities of the basilar membrane in the inner ear, which also divides up the sound into its frequency components at every point in time.

6.1 How a spectrogram is computed

Spectrograms are computed in a way that combines elements from pitch analysis techniques (section 3.1) with elements of spectral analysis techniques (section 5.1).

As with pitch analysis (section 3.1), there is the problem that a spectrum cannot be computed for a single moment in time. Instead, we have to suppose that the spectral characteristics of the sound stay constant for at least, say, 5 milliseconds. We can then cut up the sound in 5-millisecond slices and determine the spectrum of each of these slices separately. This is what is done in the following sections.

6.2 What the spectrogram of a vowel looks like

Figure 15 shows a spectrogram of the vowels [a], [i] and [u]. Time runs from left to right and frequency runs from bottom to top. The vertical stripes that we see are not the slices of section 5.1 (in good spectrographic visualizations, those are smoothed away), but the separate vocal fold vibrations. The dark horizontal bands are the formants; the harmonics of F0 cannot be seen, because 5 milliseconds is so short that such spectral detail is smeared out (that is why a spectrogram with such a short analysis window of 5 ms is called a *broadband spectrogram*).

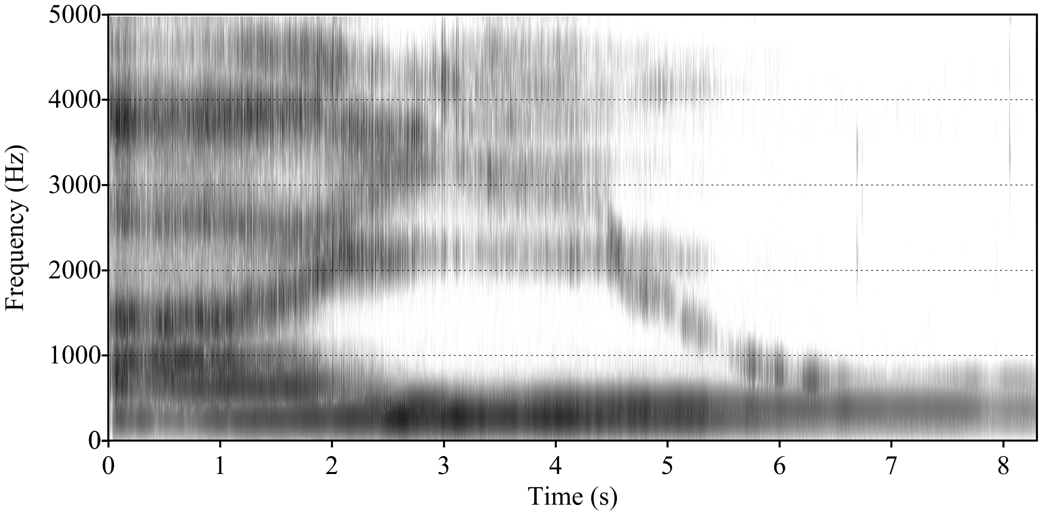

Fig. 15. Spectrogram of the vowels [a], [i] and [u].

For adult speakers, vowels can be inspected best if the visible frequency range of the spectrogram runs from 0 to 5000 Hz, because that is where the main vowel-dependent formant frequencies are. In the spectrogram of Fig. 15, the strongest frequencies are drawn in black, and the white parts of the figure depict frequencies whose strengths are 50 dB (a factor of 100,000 in power) or more below the strongest frequency in the figure; strengths in between these extremes are drawn as appropriate shades of grey. In spectrograms like these, higher frequencies are "emphasized" by 6 dB per octave with respect to lower frequencies, in order to replicate approximately how the basilar membrane integrates the power in different frequency bands; without such a "pre-emphasis", more of the upper part of the figure would have been white.

Figure 15 shows steady vowels as well as the transitions between them. The steady state of the vowel [a] is visible between 0 and 1 seconds: its first formant (F1) is given by the dark band at 700 Hz, its second formant (F2) by the band at 1400 Hz, its third formant (F3) by the band at 2600 Hz, and its fourth formant (F4) by the band at 3800 Hz. The steady state of [i] lies between 3 and 4 seconds; its F1 is seen as a strong band at 300 Hz, its F2 as a weaker band at 2300 Hz, and its F3 and F4 together form a band at 3100 Hz. How this state of affairs comes about is betrayed by the transition between 1 and 3 seconds: we see F2 rise and F3 fall, until they fall together at 2 seconds; after this, the original F2 (now by definition called F3) continues to rise above the original F3 (now by definition called F2), until it hits F4 just before 3 seconds. This crossing is repeated at the end of [i]: at 4 seconds F3 starts to fall, and it crosses F2 at 4.5 seconds (2300 Hz), thereby becoming the new F2 by definition. F2 continues to fall until it reaches 700 Hz at 6.5 seconds, whether the steady state of [u] starts. The resonance that falls from 3100 to 700 Hz corresponds to the size of the cavity in front of the oral constriction: this cavity is small for [i], which has a pre-palatal constriction, and large for [u], which has a velar constriction. For details on how formant values relate to cavities, see Fant (1960) and Stevens (1998).

It can be seen by comparing Fig. 15 with Figs. 1 to 5 that the formants of vowels are much easier to read from the spectrogram than from the waveform, not only because in the waveform the formants tend to be mingled (Fig. 4), but also because when zooming out to several seconds the formants continue to be visible in the spectrogram (Fig. 15) but not in the waveform (Fig. 5).

6.3 Other spectrograms

Figure 16 shows the spectrogram of the sound [s̢ʃɕss̻sɕʃs̢], i.e. a dynamic sibilant whose spectral centre of gravity rises from the lowest to the highest possible value and falls back again. The frequency range on the vertical axis is larger in Figure 16 than in Figure 15, since energy is concentrated at higher frequencies for fricatives than for vowels; a display depicting up to 5000 Hz only would fail to show most of the spectral energy for [ss̻s] at the centre of Figure 16.

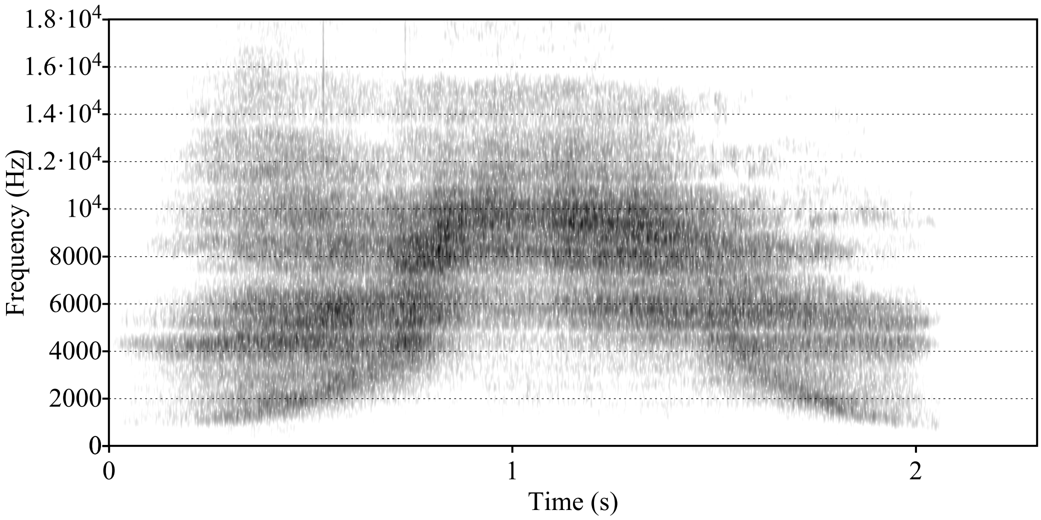

Fig. 16. Spectrogram of sibilants.

As in the case of the vowels, we see here not only the steady states, but especially the dynamic changes in the sound that result from the continuous movements of the articulators. The dynamic acoustics of the plosive of Fig. 6 and the trill of Fig. 8 are in Figs. 17 and 18, respectively.

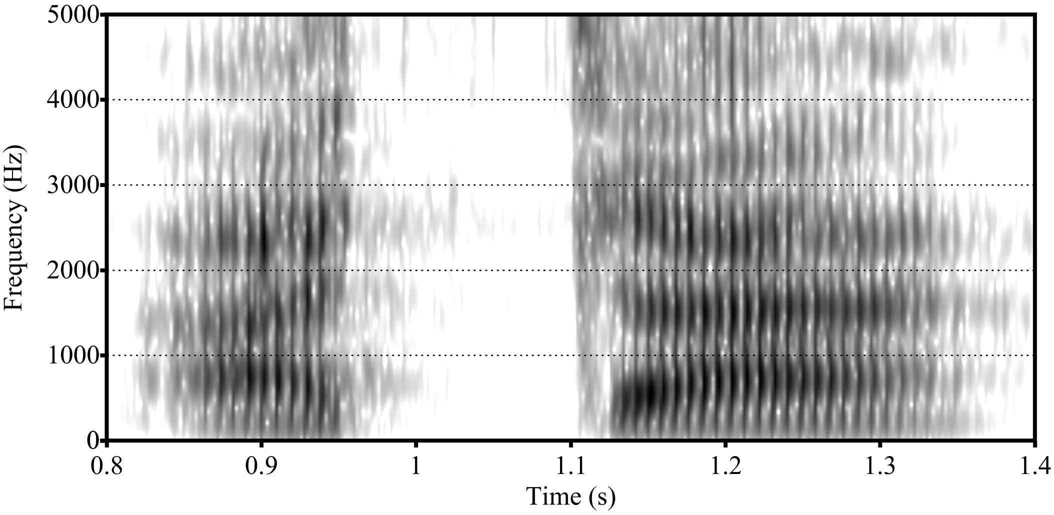

Fig. 17. Spectrogram of [aca], showing the four acoustic correlates of the plosive.

In Fig. 17 we see that the first formant of the initial [a] moves down from its steady state of 700 Hz around 0.86 seconds to almost zero at 0.95 seconds. At the same time, F2 and F3 move up and F4 approaches F3 around 3000 Hz. This closeness of F3 and F4, which is typical for palatals, can be seen even better at 1.12 seconds, where the opposite movements of all four formants can be seen. Formant transitions like these provide important cues about a consonant's place of articulation. Beside formant transitions, the plosive in Figure 17 is also characterized by the silence between 0.95 and 1.1 seconds. Such silences correspond to the closure of the active articulator (in this case, the tongue blade) against the passive articulator (the palate and upper teeth here). Plosives are also characterized by a release burst (around 1.11 seconds in Figure 17), which is caused by the sudden release (through a narrow slit) of the pressure that had built up behind the constriction during the closure.

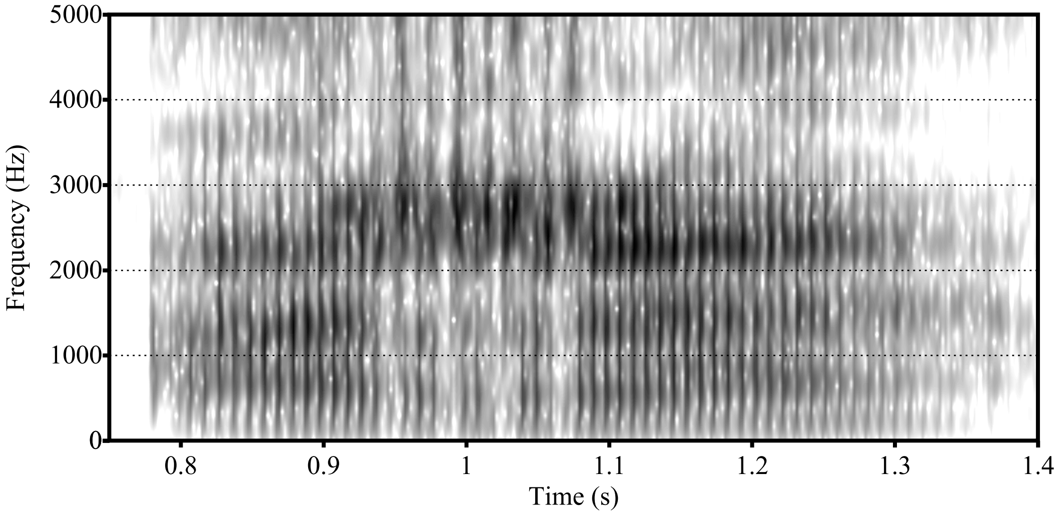

Fig. 18. Spectrogram of [ara].

In Fig. 18 we see the same four tongue-tip closures that we saw in Fig. 8, as lighter bands around 0.947, 0.985, 1.023 and 1.065 seconds. The closer vertical striping, with a period of about 0.01 second, is seen during the vowels before and after the trill and represents glottal fold vibration; this can also be seen in Fig. 17.

This section only discussed some basic aspects of the spectrograms of some speech sounds. For more of these, see the introductory textbook by Ladefoged & Disner (2012). For an in-depth treatment of the sounds of the world's languages, see the handbook by Maddieson & Ladefoged (1996). For an in-depth treatment of the causal relationships between articulation and acoustics, see Stevens (1998).

6.4 Limitations of the spectrogram

While the spectrogram visualizes the main acoustic landmarks as a function of time and frequency fairly well, it is not especially strong at visualizing the strengths of these landmarks. This is because these strengths are visualized as grey values, and the capability of the human eye to interpret more than a few different grey values at the same time is moderate. For more precise measurements one can collapse all the times of the spectrogram and obtain an average spectrum as in section 5.

7. Formant analysis

In section 6.2, I discussed how formants, the acoustic landmarks that distinguish vowel quality, can be read from the black bands in the spectrogram. Under some conditions, computer software can help automate these measurements.

7.1 Automated formant analysis techniques

Most automated formant analysis methods use a reverse all-pole filtering algorithm to extract formants (Anderson 1978). This method regards speech production as consisting of a relatively independent *source* and *filter* (Fant 1960). The "source" here is the vocal fold vibration; the "filter" consists of the vocal tract resonances, i.e. the formants, each of which can be seen as a damped sine wave. Helped by a pre-emphasis technique such as the one described here in section 6.2 to flatten the overall spectral slope, the algorithm manages to separate the source signal from the filter to some extent. The algorithm thereby manages to assign values to as many formants as you, the operator of the automated formant analysis, ask for.

7.2 Applications and limitations of the automated formant analysis

The automated analysis technique just described is notoriously brittle. It works correctly under a number of assumptions, but these assumptions can easily be violated in speech.

One assumption is that the vocal tract can be regarded acoustically as a cascade of resonating filters. This assumption is violated once there are side branches, such as those that appear in nasal consonants, lateral consonants, and nasalized vowels; such articulations cause *zeroes* in the spectrum, and these cannot be approximated well with an all-pole model. This is why automated formant analysis works best for oral vowels and glides, as in Fig. 19, and is not advisable for nasal or lateral consonants. To be sure, a side branch that is present even in oral vowels is the trachea; for [i]-like vowels it can yield a visible dark band in the spectrogram around 1500–2000 Hz (Stevens 1998: 300), which an automated measurement can incorrectly regard as an F2.

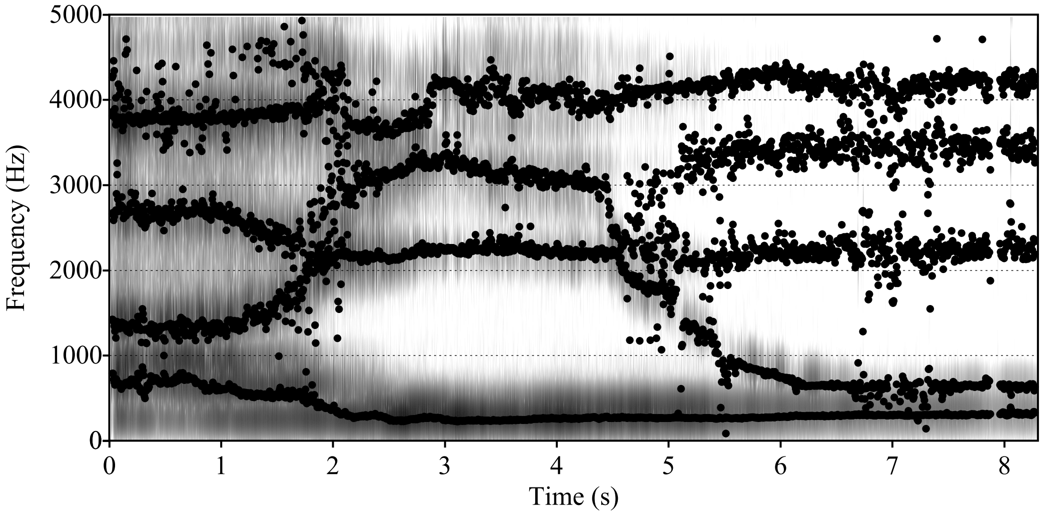

Fig. 19. Automated formant measurement in the vowels [a], [i] and [u], superimposed on the spectrogram of Fig. 15.

Another assumption is that the speech signal can be decomposed into a glottal source signal and a filter signal that represents the influence of the supraglottal vocal tract. This is not necessarily true for small children, whose very light vocal folds may become synchronized to one of the resonances, usually F1. A related requirement of the automated formant measurement method is that the spectral slope of the source is known, so that it can be compensated for; this is necessary because the reverse filtering algorithm can only be performed well on spectrally flat signals, i.e. signals where high frequencies are approximately as strong as low frequencies. As vowels on average have a 6 dB/octave falling spectral slope, i.e., high frequencies are less strong than low frequencies, the automated formant measurement method typically applies a pre-emphasis filter before performing the reverse filtering algorithm. This filter emphasizes the high frequencies by applying a 6 dB/octave rising spectral slope from 0 or 50 Hz on, which compensates exactly for the 6 dB/octave falling spectral slope. The assumption of a 6 dB/octave falling spectral slope may work for modal phonation but is violated by creaky voiced phonation and by breathy voiced phonation.

Yet another assumption is that the length of the vocal tract is known. The reverse filtering algorithm has to be told how many formants it has to find below what maximum formant frequency. For female voices it is advisable to ask for five formants between 0 and 5500 Hz, whereas for male voices the maximum formant frequency should be 5000 Hz instead. These maximum frequencies depend on the length of the vocal tract, with 5000 Hz assuming a vocal tract length of 340 metres per second (the speed of sound) divided by 5000 Hz (the maximum frequency), multiplied by 5 (the number of formants to look for), divided by 2, which makes 17 centimetres. However, using a constant number of e.g. 5500 Hz for all a person's vowels assumes that the vocal tract length is the same for all vowels spoken by that person. This assumption is generally not met: Escudero, Boersma, Rauber and Bion (2009) find that the maximum frequency for [i], a vowel that shortens the vocal tract, should be 700 Hz higher than the maximum frequency for [u], a vowel that lengthens the vocal tract.

Asking for five formants is necessary even if you are interested only in F1 and F2, because if you ask for only two formants the algorithm will distribute those two formants over the whole range from 0 to 5500 or 5000 Hz. Asking for two formants between 0 and 2200 or 2000 Hz does not work either, because F2 tends to fluctuate heavily with articulation; that is why you want to measure F2 in the first place. As the fifth and higher formants do not depend on articulation too much (except for the vocal tract shortening and lengthening effects mentioned above), and formants above the 5th or 6th may be absent from the signal because of source or recording restrictions, it is usual to ask the automated measurement method for five formants.

If after reading these cautionary lines the reader can still trust automated formant analysis for his or her applications, then he or she is invited to go ahead with it. As automated formant measurement is by far the most commonly used method for acoustically analysing vowel quality, the method can certainly make your vowel quality data comparable with data published by others. To enhance reliability, you are advised to take multiple measurements and then to take the median of the measured values, so that the influence of gross measurement errors is minimized.

8. Conclusion

This chapter explained how you can measure acoustic properties of speech signals by hand. This is certainly a feasible line of approach if the number of sounds to be measured is limited. For larger datasets, acoustic analysis can be automated by annotating landmarks in the acoustic signal (e.g., the start and end points of a vowel) and using “scripts” in the analysis software to extract the needed acoustic measures (e.g., the first and second formant at the vowel midpoint). A discussion of such procedures lies beyond the scope of this chapter, but the internet provides many resources for this purpose, including tutorials on how to write scripts, as well as existing scripts that can be modified to obtain the measures needed for your specific project.

This short chapter has not been able to explain everything there is to know about acoustic measurements. For a very readable introduction, see Ladefoged and Disner (2012). For a handbook on the sounds of the world’s languages, see Ladefoged and Maddieson (1996). For a technical overview, see Stevens (1998). There is also a wealth of literature on detailed acoustic correlates of many speech sounds and prosodic structures.

References

Anderson, N. (1978). On the calculation of filter coefficients for maximum entropy spectral analysis. In Childers: *Modern spectrum analysis*. IEEE Press. 252–255.

Boersma, Paul (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences (University of Amsterdam)* 17: 97–110.

Boersma, Paul (2009). Should jitter be measured by peak picking or by waveform matching? *Folia Phoniatrica et Logopaedica* 61: 305–308.

Boersma, Paul, & David Weenink (1992–2012). Praat: doing phonetics by computer. Computer program available from http://www.praat.org.

Deliyski, Dimitar D., Heather S. Shaw, & Maegan K. Evans (2005). Adverse effects of environmental noise on acoustic voice quality measurements. *Journal of Voice* 19: 15–28.

Escudero, Paola, Paul Boersma, Andréia S. Rauber, & Ricardo A.H. Bion (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *Journal of the Acoustical Society of America* **126**: 1379–1393.

Fant, Gunnar (1960): *Acoustic theory of speech production*. Mouton, The Hague.

Fourier, Jean Baptiste Joseph (1822). *Théorie analytique de la chaleur*. Paris: Chez Firmin Didot, père et fils.

Ladefoged, Peter, & Sandra Disner (2012). *Vowels and consonants*. 3rd edition. Malden & Oxford: Blackwell.

Ladefoged, Peter, & Ian Maddieson (1996). *The sounds of the world’s languages*. Malden & Oxford: Blackwell.

Stevens, Kenneth N. (1998). *Acoustic phonetics*. Cambridge, Mass. & London: MIT Press.

Talkin, David T. (1995). A robust algorithm for pitch tracking. In W. Bastiaan Kleijn & Kuldip K. Paliwal (eds.) *Speech coding and synthesis*. Amsterdam: Elsevier. 495–518.