9. OPTIMIZING THE EXTRACTION OF VOWEL FORMANTS[1]

Radek SKARNITZL – Jitka VAŇKOVÁ – Tomáš BOŘIL

9.1. Introduction

Vowel formants have been scientifically analyzed for at least 150 years – since 1863, when Hermann von Helmholtz devised special resonators to imitate the resonances of vocalic sounds – and they continue to be the most frequently applied parameter for the description of vowels in languages of the world. Formants are studied when we are interested in the vowels of a specific language (e.g., Hillenbrand, Getty, Clark & Wheeler, 1995 for American English; Deterding, 1997 for Standard British English; Beňuš, 2012 for Slovak; Skarnitzl & Volín, 2012 for Czech) or when we are comparing different languages or language varieties (e.g., Escudero, Boersma, Rauber & Bion, 2009 for Brazilian and European Portuguese; Fox & Jacewicz, 2009 for American English; Ferragne & Pellegrino, 2010 for British English), when we are analyzing sound change or socially conditioned varieties (e.g., Labov, 1963; Fabricius, 2002, 2007; de Jong, McDougall & Nolan, 2007; Harrington, Kleber & Reubold, 2007), or when looking at manifestations of the speaker's individuality (e.g., McDougall, 2006; Moos, 2012; or Fejlová, Lukeš & Skarnitzl, 2013 for Czech).

The appeal of vowel formants for speech scientists is understandable, given their intuitive character and transparency on multiple levels. Using only two values – that of F1 and F2 – we are typically able to distinguish all vowels of a language or a language variety. Moreover, when plotted in a conventional two-dimensional chart, formant values correspond to the articulatory settings of the tongue.

1 The authors would like to thank Alžběta Růžičková for her help with the manual labelling of vowel formants. This research was supported by the project GAČR 406/12/0298, by the Internal grants 2014 VG184 solved at the Faculty of Arts in Prague granted to the second author, and by the Programme of Scientific Areas Development at Charles University in Prague (PRVOUK), subsection 10 – Linguistics: Social Group Variation.

Given their transparency and widespread use, then, it might seem misguided to relate vowel formants to complexity. However, while it is true that using formants for describing vowels is elegant and straightforward, formant analysis itself represents an inherently complex endeavour. This chapter thus addresses formant extraction – something that phoneticians do on a regular basis but rarely give this process much thought, because formant values are just a press of a button away in the most frequently used software packages for phonetic analysis. Although these software packages like Praat (Boersma & Weenink, 2014) or WaveSurfer (Sjölander & Beskow, 2005) do recommend "default" settings for formant extraction, which are clearly based on average vocal tract size of male and female speakers, these settings have not, to our best knowledge, been subjected to empirical examination. The aim of this chapter is therefore twofold. First, to compare a number of extraction settings in Praat and Snack (Sjölander, 2014; the implementation of formant analysis in WaveSurfer) and to determine whether the default settings really do perform the best. The second aim is to compare the performance of Praat and Snack, since formant analysis as implemented in Praat has frequently been criticized as considerably worse than that in WaveSurfer.

9.2. Stipulating "ground truth"

Any attempts to compare the performance of algorithms and their settings require a reference set of manually-labelled formant values, what has been called the "ground truth" by Deng, Cui, Pruvenok et al. (2006). Deng and his colleagues, in what to our knowledge is the only reference set of this kind, selected 538 utterances from the TIMIT database (Garofolo et al., 1993) and labelled F1–F3 in every 10-ms frame; in other words, they were interested in vocal tract resonances in all speech sounds, not only in vowels. Unfortunately, Deng et al. (2006) do not provide much detail on how the manual measurements were conducted, especially what (if any) were the instructions given to the labellers. The labellers were simply asked to click with the mouse in the spectrogram where they believed the resonance to be located. The authors did test between-labeller variation on a subset of the utterances and reported the following absolute deviations for vowel segments: 55 Hz for F1, 69 Hz for F2, and 84 Hz for F3. Although they mention that this was higher than expected, they did not follow up and give more detailed instructions to the labellers. Therefore, there was no attempt to resolve the inevitable inter-labeller discrepancies. Duckworth, McDougall, de Jong and Shockey (2011) did provide their labellers with a set of instructions on formant measurement, but they still seem insufficient for our purposes. The aim of our

research was to prepare a more modest database of hand-labelled vowel formant values in Czech, but with more explicit guidelines for the labellers which included stipulating the "ground truth" for a given formant based on more than a single labeller.

9.2.1. Method

To obtain our "ground truth", we analyzed 5 tokens of each of the 5 Czech short monophthongs /ɪ ɛ a o u/ from 10 native speakers of Czech (5 females, 5 males). In total, then, we worked with 250 vowel tokens. The values of vowel formants (F1–F3) were visually determined in the midpoint of each vowel. The vowels were analyzed in Praat by two pairs of labellers; in other words, F1–F3 for every vowel were identified by two labellers. The complete dataset from this stage thus comprised 1,500 readings (250 vowels × 3 formants × 2 labellers).

The guidelines given to the labellers concerned two areas: viewing the speech signal (spectrogram) in Praat, and identifying the formant values. With respect to the first area, the Praat Edit window was always maximized on the computer screen, the viewing range was set to 0–3 kHz for male voices and 0–3.5 kHz for female voices. The most important setting was the length of the analysis window: we opted for a window length of 10 ms, between the traditional wideband spectrogram with a 5-ms window and narrowband spectrogram with a 30-ms window. The reading of formant values was to be based primarily on information available in the spectrogram – by default, the formant contours in Praat were not displayed, and the instruction could be summarized as "keep clicking until the cursor is visually located in the centre of the formant". When a formant was not clearly visible or when there were competing spectral peaks, the labellers were allowed to turn on the display of formant contours in Praat (the default settings were used for female and male speakers), or to visualize the FFT spectrum of the target vowel. All these possibilities are illustrated in Figure 9–1.

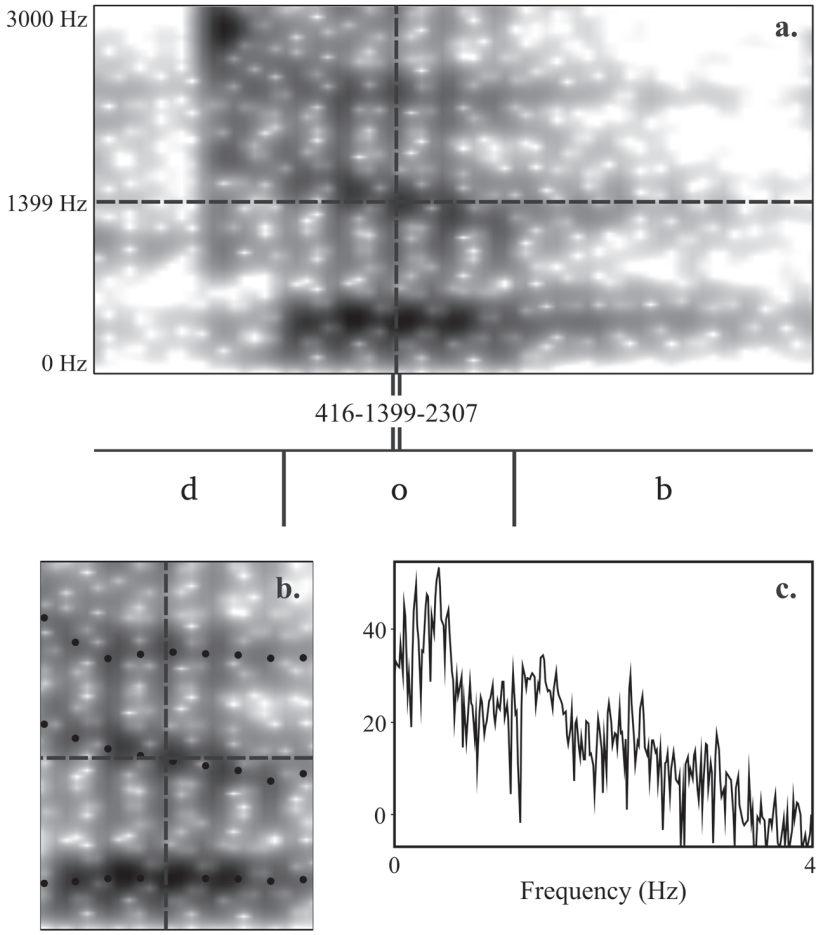

Figure 9–1: Illustration of the alternatives for reading formant values: a. the default option based only on spectrogram information; b. with formant contours visualized; c. using the FFT spectrum of the target vowel.

9.2.2. Results

The first step was to simply compare the manual measurements of the two labellers. In Table 9–1, we can see that mean absolute deviations are comparable with those reported by Deng et al. (2006), only that of F2 is considerably lower in our data. The table also indicates how large the deviations were. Out of the 750 measurements, the two labellers deviated by less than 30 Hz in 583 cases; on the other hand, the deviation exceeded 150 Hz in 23 cases, most frequently for F3.

	MAD	> 30 Hz	> 50 Hz	> 100 Hz	> 150 Hz
F1	50.8 Hz	46	22	8	3
F2	49.3 Hz	57	39	14	5
F3	83.6 Hz	64	46	24	15

Table 9–1: Mean absolute difference (MAD) of formant values between the two labellers, and the number of deviations exceeding 30, 50, 100 and 150 Hz.

We decided to analyze the 46 items where the labellers differed in their readings by more than 100 Hz together and tried to reach consensus. It transpired that in some of the cases (especially those where the difference concerned F1 or F2) one of the labellers misidentified a vowel formant with a nasal formant. The analyzed vowels in our research did not appear before nasal consonants, but some did appear after a nasal consonant. We did not expect the effect of progressive nasalization to persist in the midpoint of the post-nasal vowel but in several cases it did. The situation was more complicated in some of the F3 items but, in the end, consensus was reached for most (but not all) items. Table 9–2 shows mean deviations and the magnitude of the deviations after the corrections have been made.

	MAD	> 30 Hz	> 50 Hz	> 100 Hz	> 150 Hz
F1	29.0 Hz	37	12	0	0
F2	29.1 Hz	47	25	0	0
F3	32.9 Hz	55	26	1	1

Table 9–2: Mean absolute difference (MAD) of formant values between the two labellers, and the number of deviations exceeding 30, 50, 100 and 150 Hz; after corrections (see text).

A mean absolute difference of around 30 Hz was considered a more acceptable deviation – it is far superior to the results reported by Duckworth et al. (2011), for instance – and we could therefore proceed with the final step of stipulating the "ground truth" formant values, which were calculated by simply averaging the pair of corrected values from the two labellers. These are the values which were subsequently compared with the automatically extracted values in Praat and Snack under different settings. However, before we turn to this comparison, we would like to continue discussing manual formant measurement from a different, and very important perspective.

9.3. Uncertainty of "ground truth"

Every measurement is affected by some degree of uncertainty. In this section, we try to uncover the most important causes of random biases influencing the final measured value, and estimate the size of their effect. Let us focus on the following problems connected with manually determined formant values from the spectrogram:

- resolution of the screen and mouse,
- spectrogram resolution,
- estimation of formant frequency as the darkest line produced by smoothing of harmonics,
- between-labeller variability.

9.3.1. Screen and mouse resolution

Although the frequency cursor in the spectrogram shows values with round off to units of hertz (see Figure 9–1a), the frequency range of 0–3 or 3.5 kHz occupies only a small fraction of the screen resolution, and therefore the screen and mouse resolution limits the possible values to discrete states with a much larger spacing. A sample histogram of obtained values by one labeller is depicted in Figure 9–2; it would not be possible, for example, to "measure" 400 Hz. The steps between the bins are not identical, but the approximate distance is 24 Hz, which we can assume to be a firm estimate of the resolution of the values caused by the screen and mouse.

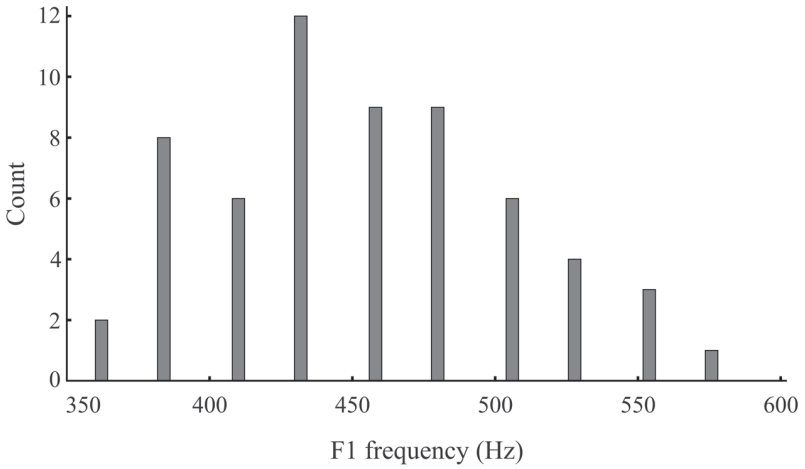

Figure 9–2: A sample histogram of all female F1 values below 600 Hz manually measured by one labeller. The distribution is discrete with the step not being equal but approximately 24 Hz.

9.3.2. Spectrogram resolution

After addressing the question as to how precisely one can read a frequency value from the screen, the very source of the displayed information, i.e. the spectrogram resolution, should also be investigated. The spectral domain always involves a compromise between time and frequency resolution: the rougher the time resolution, the better the frequency resolution. With our segment of 10 ms (*cf.* section 9.2.1), the standard distance between frequency bins in an FFT spectrum is $\Delta f = 1 / 10\ \text{ms} = 100\ \text{Hz}$. The process of spectrogram computation uses the trick called zero padding, which artificially increases the duration of a segment by appending extra zero samples. Zero padding does not allow any additional useful information to be extracted from the measured signal, but it does result in a finer spectral resolution; however, in fact, the impact of this operation leads to the uncovering of the windowing effect known as spectral leakage. Each harmonic in the original signal convolves with the spectrum of the segmentation window, which consists of the main lobe (centred at the position of the original harmonic) and sidelobes. Since the speech signal contains a lot of harmonics, the resulting spectrum is a mixture of these convolved components. The default rectangular window (see Figure 9–3a) has a relative width of the main lobe equal to Δf, which is the best possible value among window functions, but it has the worst peak level of the sidelobes (Oppenheim, Schafer & Buck, 1999). To attenuate the sidelobes, other windows with higher relative width of the main lobe (such as Praat default Gaussian window; Harris, 1978) are used in spectral analysis (see Figure 9–3 b). For that reason, even with a finer resolution of the FFT of a zero-padded segment, two harmonics which are closer together than Δf cannot be correctly discerned.

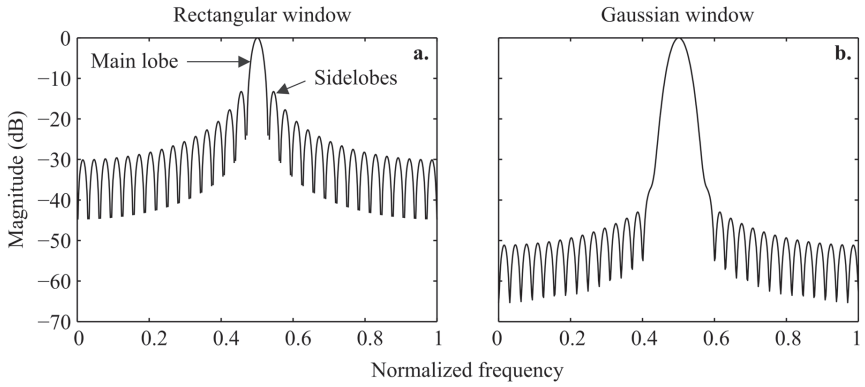

Figure 9–3: Main lobe and sidelobes of a. Rectangular window, b. Gaussian window.

However, even for harmonics with larger values, distortion of the spectrum caused by the summation of the main lobe together with the sidelobes of other harmonics still exists, leading to an inaccurate measurement of spectral peak positions, as illustrated in Figure 9–4. Fortunately, formant measurement does not need the display of all the harmonics; rather, what we are analyzing is spectral envelope peaks. Still, the uncertainty of frequency peak measurement in a short signal segment should be taken into account. It is difficult to estimate the resolution of the mean spectral envelope position because it ranges from a few hertz to one hundred hertz in the four examples depicted in the figure, depending on fo. For the purpose of this chapter, let us assume the spectrogram resolution to be approximately 30 Hz.

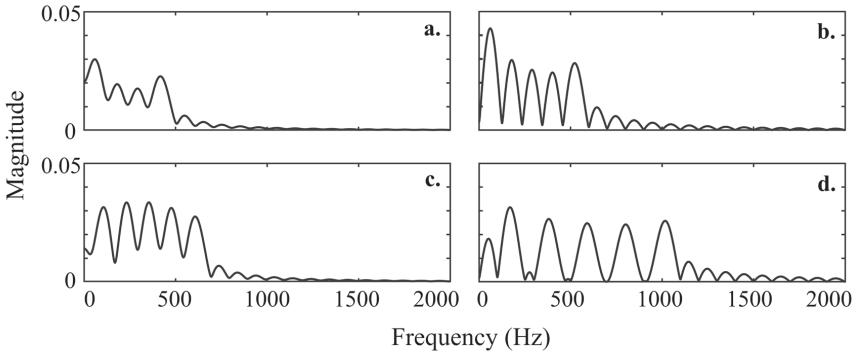

Figure 9–4: Spectral leakage effect mixing main lobes with sidelobes leads to inaccurate spectral peaks position. A 10-ms signal segment multiplied by the Gaussian window (Praat default) padded with extra 625 ms of zero samples leading to frequency resolution of 1 / 635 ms = 1.57 Hz. All four examples consist of an artificial signal with five sine waves simulating fo and four additional harmonics (with integer multiples of fo frequency). a. fo = 80 Hz, main lobes are so close together that only four peaks are present in the spectrum (58 Hz, 181 Hz, 292 Hz, 417 Hz). b. fo = 100 Hz, five main peaks (61 Hz, 178 Hz, 289 Hz, 401 Hz, 523 Hz). c. fo = 120 Hz, five main peaks (106 Hz, 232 Hz, 354 Hz, 477 Hz, 607 Hz). d. fo = 200 Hz, six main peaks (52 Hz, 169 Hz, 381 Hz, 591 Hz, 801 Hz, 1017 Hz), the first peak is an artefact of the summation of sidelobes.

6.3.3. Formant frequency estimation

When stipulating our ground truth, that is, manually determined formant values, formant frequencies cannot be measured directly. A formant can only be estimated from the speech signal (which is the convolution of glottal pulses and the vocal tract response) as the dark line in the spectrogram (i.e., a peak of an average envelope created by harmonics). Although the resonance frequency of the vocal tract (the formant frequency) may be constant, the position of the dark line will change even with small changes of fo. We have modelled this situation using

PSOLA (pitch-synchronized overlap and add, Moulines & Charpentier, 1990) algorithm in Praat where fo frequency was manipulated, while formant frequencies were kept constant. Then, two labellers were asked to estimate the frequency of F1 and F2 from the signal. Both agreed on the same values, with the maximum difference in the visual estimates being around 10 Hz for F1 and 30 Hz for F2. However, the formant values themselves shifted by 80 Hz, both for F1 and F2, depending on the magnitude of change of fo. We have to realize that formant values should, theoretically, stay the same and not be affected by fo changes. This means that we can assume the uncertainty component resulting from estimating formants from the spectrogram to be approximately 80 Hz.

9.3.4. Uncertainty factors combined

After between-labeller variation was reduced by the consensus corrections mentioned in section 9.2.2, the mean difference between two labellers was approximately 30 Hz for all formants (F1, F2, and F3). This resolution constitutes the fourth component of the uncertainty estimate.

All these four types of uncertainty can be marked as "type B" evaluations – uncertainty estimated from information other than using repeated readings and statistics, i.e. estimated from past experience, calibration, calculations, common sense etc. (Bell, 2001). Assuming a uniform distribution of these four factors, the estimate of an expanded combined standard uncertainty using the coverage factor 2 to give a 95% confidence level is:

$$2\times\sqrt{\left(\frac{24/2}{\sqrt{3}}\right)^2+\left(\frac{30/2}{\sqrt{3}}\right)^2+\left(\frac{80/2}{\sqrt{3}}\right)^2+\left(\frac{30/2}{\sqrt{3}}\right)^2}=\pm 54\ \text{Hz}.$$

This estimate, based on the inaccuracies assumed in sections 9.3.1–9.3.3, should be regarded only as a rough indication of the uncertainty of manual formant measurement from the spectrogram. When performing manual formant measurement, one must therefore keep in mind that the actual formant values may lie in the range of approximately ±54 Hz around the manually determined value at the 0.05 level of significance.

9.4. Performance of Praat and Snack

Having determined the "ground truth" and discussed the inaccuracies inevitably related to it, our next aim was to compare the manually measured formant values with the performance of two software tools which extract formant frequencies automatically – Praat and Snack. Both Praat

and Snack use linear predictive coding (LPC) to locate formants, but employ different methods of estimating the coefficients. In order to find an optimal setting, i.e. one whose output will be closest to our ground truth, we manipulated two parameters. First, it was the order of LPC, which determines the number of formants (*n*) which are detected, with the order equal to 2*n*. The second varied parameter was the maximum frequency at which the *n*-th formant could be detected. We did not change the pre-emphasis value from that set as default (but see Harrison, 2004). We will present results for Praat first and then compare them with those for Snack.

9.4.1. Praat measurements

The default settings for formant extraction in Praat order are 10th order LPC, which means that 5 formants are detected, in the 0–5 kHz range in male voices and in the 0–5.5 kHz range in female voices. We varied LPC order between 6 and 12, with the maximum frequency of a formant being set to 5,000 Hz or 5,500 Hz, according to the above-mentioned default settings for male and female speakers. For each vowel token, we thus tested 14 different settings in total: LPC6 – 3000 Hz, LPC6 – 3300 Hz, LPC7 – 3000 Hz, LPC7 – 3300 Hz, LPC8 – 5000 Hz, LPC8 – 5500 Hz, LPC9 – 5000 Hz, LPC9 – 5500 Hz, LPC10 – 5000 Hz, LPC10 – 5500 Hz, LPC11 – 5000 Hz, LPC11 – 5500 Hz, LPC12 – 5000 Hz, and LPC12 – 5500 Hz. Since the first four settings, where the maximum frequency of a formant was either 3000 or 3300 Hz, frequently yielded undefined values (especially of F3), these settings were disregarded from further analyses. The subsequent comparisons are therefore based on the remaining 10 settings (LPC orders 8–12, formant detection in the 0–5 or 0–5.5 kHz range).

The performance of Praat, separately for male and female speakers, for all three formants (F1–F3) is illustrated in Figure 9–5, which assesses it by means of deviation (in Hz) of the automatically measured value from the manually determined ground truth for all 10 settings mentioned above. Taking a look at the whole picture, it is clear that there is huge variability in how the individual settings perform: for example, deviations for F3 range from below 100 Hz to 800 Hz. The result of the default setting for the given gender is always marked with slanted lines, and we can see that it performs considerably well in all cases. Yet, a closer look reveals that there are settings which perform systematically, albeit slightly better than the default one, namely LPC9 in the 0–5 kHz range for female speakers and LPC11 in the 0–5.5 kHz range for male speakers.

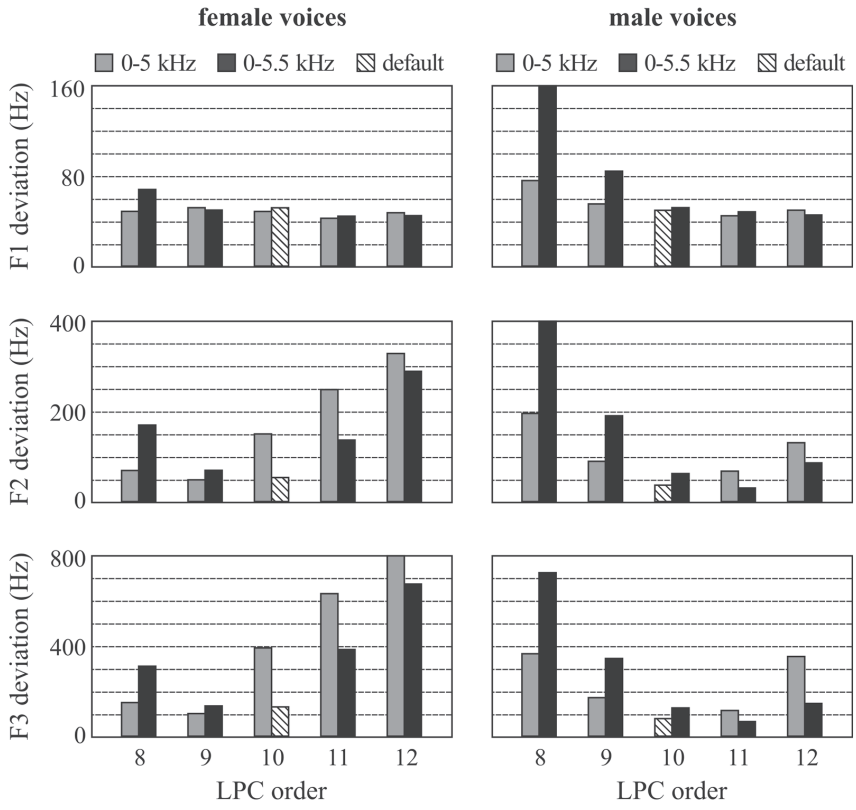

Figure 9–5: Deviations (in Hz) of automatically measured formant values (F1–F3) in Praat from ground truth for male and female speakers, using 10 different formant extraction settings (LPC order 8–12, maximum frequency 5 or 5.5 kHz). The default setting is marked by slanted lines.

In the case of F1, the improvement is, in fact, negligible: the two settings yield the same deviation from ground truth for female speakers, and the difference is merely 2 Hz for male speakers. The differences between the two settings are similarly small also for F2 (6 and 4 Hz for female and male speakers, respectively). The default setting is outperformed to a greatest extent in the case of F3, where the improvement comprises 29 Hz for female speakers and 13 Hz for male speakers. However, considering the uncertainties of manual measurement discussed in section 9.3, the improvement of our "superior" settings over the default settings is in fact negligible.

So far, we have been presenting the results in terms of the mean deviation of the extracted values from ground truth, primarily due to its intuitive character. Nevertheless, it is preferable, when comparing the performance of algorithms, to use root-mean-square deviation (RMSD)

instead of mean deviation, as the RMSD represents the sample standard deviation of the differences. The mean deviation averages all deviations and larger differences thus get blurred. In contrast, RMSD "penalizes" such outliers more, and they become more visible. A comparison of the differences between ground truth on the one hand and the default and the slightly superior settings on the other hand as expressed by mean deviation (MD) and root-mean-square deviation (RMSD) is presented, for male speakers only, in Table 9–3.

MALES	F1		F2		F3	
	MD	**RMSD**	**MD**	**RMSD**	**MD**	**RMSD**
LPC10-5kHz	50	69	39	58	81	142
LPC11-5.5kHz	48	65	35	48	68	107

Table 9–3: Comparison of the differences observed between ground truth and the default setting as well as ground truth and the slightly superior setting in Praat, expressed as mean deviation (MD) and root-mean-square deviation (RMSD) for F1–F3 of male speakers.

The table shows, on the one hand, that the two settings are largely comparable and, on the other hand, that the RMSD values are in all cases higher than the MD values (by the virtue of penalizing the outliers), and this difference is most pronounced for F3 measurements. Since RMSD reflects the performance of an algorithm more reliably, it will be employed also in the following evaluations.

9.4.2. Snack measurements

The second software tool which was tested in this study was Snack, which is implemented in the well-known WaveSurfer or more recently also in VoiceSauce (Shue, 2013). We can conclude from a number of informal discussions among phoneticians from various countries that it is commonly believed that formant analysis performed by Snack is superior to the one implemented in Praat. We were therefore especially interested in the comparison of our ground truth with automatically extracted formant values in Snack.

In total, 12 different settings were tested: both for autocorrelation and stabilized covariance, the LPC order was varied from 10 to 12, the maximum frequency being set to 5 or 5.5 kHz.

The setting of Snack which performed the best results, i.e. closest to our ground truth as quantified by RMSD, was the default autocorrelation, using LPC12 and the maximum frequency for formant detection being 5 kHz (lower LPC orders performed significantly worse). In order to assess the performance of Praat and Snack, this setting was compared with the default setting in Praat (for male and female speakers separately), as

well as with the slightly superior setting discussed in the previous section. The deviations of the three methods from ground truth are plotted in Figure 9–6.

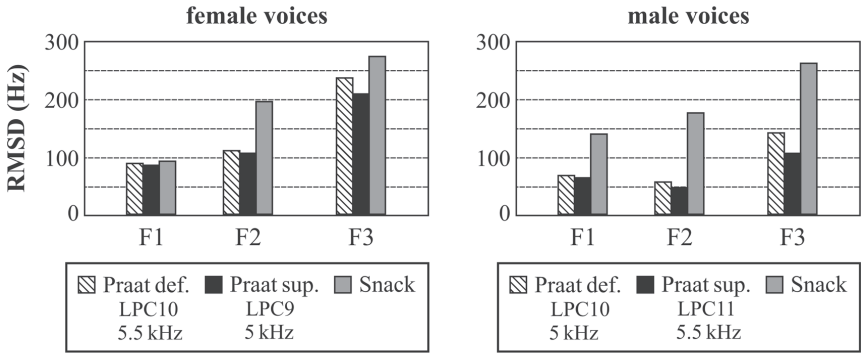

Figure 9–6: Root-mean-square deviations of automatic formant measurements (F1–F3) by the Praat default setting, a slightly superior Praat setting, and Snack for male and female speakers.

The chart revealed that Snack was considerably worse in all instances, which was rather unexpected given the fact that Snack is, as already mentioned, generally considered to be superior in formant estimation than Praat.

There is another factor worth mentioning in relation to formant analysis in Snack. Table 9–4 shows the comparison of four settings in Snack with ground truth. We can see that extraction in the 0–5.5 kHz range is significantly worse than in the 0–5 kHz range, especially for male speakers. It thus appears that it is the autocorrelation *vs.* covariance setting which differentiates the most successful setting for male and female speakers, with autocorrelation performing best in females and covariance in males.

	FEMALES			MALES		
	F1	**F2**	**F3**	**F1**	**F2**	**F3**
LPC12 – 5 kHz – autocorrelation	94	197	274	162	260	384
LPC12 – 5.5 kHz – autocorrelation	147	194	314	180	427	416
LPC12 – 5 kHz – covariance	134	219	303	140	177	263
LPC12 – 5.5kHz – covariance	154	249	315	131	243	406

Table 9–4: Root-mean-square deviation (in Hz) of several extraction settings in Snack from manually stipulated ground truth for female and male speakers.

9.4.3. Praat tracker

The result of the comparison of Praat and Snack is all the more surprising considering their differences in the nature of formant estimation – while the default method of formant analysis in Praat is formant extraction, Snack employs formant tracking. The process of formant extraction involves a simple search for the most probably value of a formant in each consecutive frame. Theoretically, then, values of neighbouring formants might yield smaller or larger jumps, i.e. they might not be "neighbouring" (in terms of frequency) at all. In contrast, formant tracking takes context into account: it can be described as finding the cheapest path through neighbouring formant values. In other words, formant tracking tries to prevent large frequency jumps from one frame to the next, which is typically done by means of the Viterbi algorithm (Viterbi, 1967).

Though the default – and recommended – way of estimating formant frequencies in Praat is formant extraction as mentioned above, there *is* a tracker implemented in Praat. Since it is suggested by the authors in Praat manual that formant tracking should be used only for vowel and vocoids, our aim was to examine whether formant tracks can be used on our material so that the performance of Praat (extractor), Snack (tracker) and Praat (tracker) can be assessed. The Praat tracker requires reference formant values, which have been specified as odd multiples of 500 Hz for male speakers and of 550 Hz for female speakers (i.e., 550 Hz for F1, 1650 Hz for F2 etc.). As the input, the tracker typically uses formant estimates for five formants, from which it extracts three tracks (i.e., the track of F1–F3). Obviously, this poses limitations on the settings we could vary in our experiment: we examined the same six settings as in the case of Snack, i.e. LPC order 10–12 and the maximum frequency of a formant 5 or 5.5 kHz. The setting performing best – that is, showing the lowest RMSD from ground truth – was then compared with the best setting in Praat extractor (Snack was omitted from the comparison, as it was outperformed by Praat).

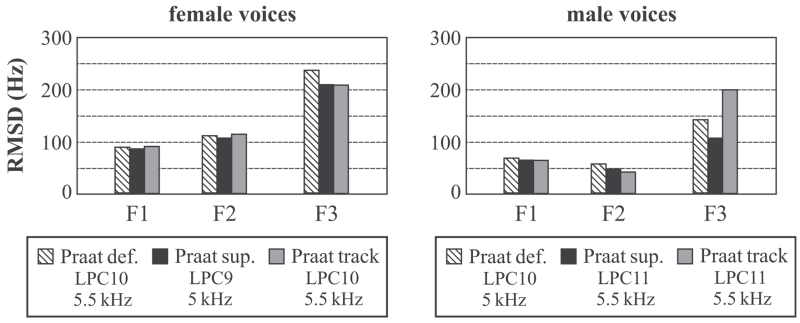

Figure 9–7: Root-mean-square deviation (in Hz) of a Praat superior setting (LPC9, maximum frequency 5 kHz) and the best Praat tracker setting from manually stipulated ground truth for female and male speakers.

The outcome of the comparison is illustrated in Figure 9–7, which shows that the best setting of the Praat tracker is comparable with our superior setting of the Praat extractor with the exception of F3 in male speakers, where the tracker performs significantly worse. The reason for this discrepancy is not clear.

9.5. Discussion and conclusion

This chapter addressed the complexities of vowel formant extraction. Although extracting formant values is a simple enough task in the most frequently used software packages for phonetic analyses, in the sense that just a few clicks of the mouse will produce the requested values, we have shown in this chapter that formant analysis is, indeed, a considerably complex endeavour. In the introduction, we mentioned the transparent relationship between vowel articulation and acoustics: when depicted in the vowel quadrilateral, the tongue positions correspond to formant values. However, the result of our experiment presented in section 9.3.3, where manipulations of fo led to markedly different estimates of vowel formants, shows that even this articulatory-to-acoustic mapping – the position of the tongue and the lips and the corresponding resonance frequencies of the vocal tract – is more complex than one might assume. Simply said, when one "digs deeper", the complexity of speech manifests itself even in seemingly straightforward relationships.

The objectives of this chapter were to test the performance of default settings in Praat and Snack *vis-à-vis* a number of settings which are possible (and feasible) in these programmes, and to compare the performance of these two analysis tools with each other. The first step consisted in the manual identification of formant values; although this "ground truth" is inherently associated with measurement uncertainty (see section 9.3), our consistency was much higher than that reported by Duckworth et al. (2011), possibly thanks to the more detailed instructions given to the labellers. The following step was to compare the manually determined and automatically extracted formant values. Surprisingly – at least given the frequent criticism of formant analysis in Praat – Praat turned out to be significantly superior to Snack, especially in the case of F2 for both genders. One may argue that this result may be due to the fact that the manual "ground truth" measurements were performed in Praat, and hence the deviation of Praat is predictably lower than that of Snack. However, let us repeat that a great majority of the ground truth values were obtained simply by visual identification in the spectrogram, without assistance from the Praat formant extractor. This could therefore not have given Praat an "edge" in the subsequent comparisons; the spectrogram

would have been the same in any software tool. We also used the formant tracker embedded in Praat and compared the results with ground truth; however, the application of the tracker did not lead to consistent improvements.

Some researchers have also suggested that it might be advisable to use vowel-specific settings for formant extraction, since spectra of individual vowel qualities look very differently. Harrison and Clermont (2012) showed that automatic formant estimation in Praat is more accurate with a different LPC order for close, mid and open vowels: the default setting (LPC10) turned out to be most successful for mid vowels like [e o], while the 12th-order LPC was best for close vowels like [i u] and the 8th-order LPC for open vowels like [a]. Indeed, given the spectral differences between the vowels, this seems to be intuitive. Nevertheless, the results of our analyses (formant extraction in Praat) do not support this appealing hypothesis: vowel-specific settings such as those reported by Harrison and Clermont (2012) did not lead to lower deviations from the ground truth values in our data. Based on the analyses of our limited dataset, we may therefore recommend the following settings for formant analysis in Praat, regardless of vowel quality: 9th-order LPC with extraction in the 0–5 kHz range for female speakers (i.e., the number of formants to be detected is set to 4.5 in Praat), and 11th-order LPC in the 0–5.5 kHz range for male speakers (the number of formants is set to 5.5). However, given the uncertainty associated with the manual measurement, the difference approaches significance only in the case of F3.

We would like to raise one more point. It must be kept in mind that our analyses were based on high-quality studio recordings, although the speech itself was comparatively natural. It is possible, and our preliminary analyses appear to support this hypothesis, that the performance of Snack might improve relative to that of Praat in recordings which feature degraded acoustic conditions, for instance background noise or mobile phone transmission.

References

Bell, S. (2001), A beginner's guide to uncertainty of measurement, *Measurement Good Practice Guide*, 11.

Beňuš, Š. (2012), Phonetic variation in Slovak yer and non-yer vowels, *Journal of Phonetics*, *40*, pp. 535–549.

Boersma, P. – Weenink, D. (2014), *Praat – Doing phonetics by computer* (Version 5.4), Retrieved on October 14, 2014 from http://www.praat.org.

De Jong, G. – McDougall, K. – Nolan, F. (2007), Sound Change and Speaker Identity: An Acoustic Study. In: Müller, Ch. (ed.), *Speaker Characteristics II*, pp. 130–141, Berlin: Springer Verlag.

Deng, L. – Cui, X. – Pruvenok, R. – Huang, J. – Momen, S. – Chen, Y. – Alwan, A. (2006), A database of vocal tract resonance trajectories for research in speech processing. In: *Proceedings of ICASSP 2012*, pp. 60–63.

Deterding, D. (1997), The Formants of Monophthong Vowels in Standard Southern British English Pronunciation, *Journal of the International Phonetic Association*, *27*, pp. 47–55.

Duckworth, M. – McDougall, K. – de Jong, G. – Shockey, L. (2011), Improving the consistency of formant measurement, *International Journal of Speech, Language and the Law*, *18*, pp. 35–51.

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

Fabricius, A. (2002), Weak vowels in modern RP: An acoustic study of happy-tensing and kit/schwa shift, *Language Variation and Change*, *14*, pp. 211–237.

Fabricius, A. (2007), Vowel formants and angle measurements in diachronic sociophonetic studies: foot-fronting in RP. In: *Proceedings of 16th ICPhS*, pp. 1477–1480.

Fejlová, D. – Lukeš, D. – Skarnitzl, R. (2013), Formant Contours in Czech Vowels: Speaker-discriminating Potential. In: *Proceedings of Interspeech 2013*, pp. 3182–3186.

Ferragne, E. – Pellegrino, F. (2010), Formant frequencies of vowels in 13 accents of the British Isles, *Journal of the International Phonetic Association*, *40*, pp. 1–34.

Fox, R. A. – Jacewicz, E. (2009), Cross-dialectal variation in formant dynamics of American English vowels, *Journal of the Acoustical Society of America*, *126*, pp. 2603–2618.

Garofolo, J. et al. (1993), *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*, Philadelphia: Linguistic Data Consortium.

Harrington, J. – Kleber, F. – Reubold, U. (2007), /U/-fronting in RP: A link between sound change and diminished perceptual compensation for coarticulation? In: *Proceedings of the 16th ICPhS*, pp. 1473–1476.
Harris, F. J. (1978), On the use of windows for harmonic analysis with the discrete Fourier transform. In: *Proceedings of the IEEE*, *66*, pp. 51–83.

Harrison, P. (2004), *Variability of formant measurements*, York: University of York, (MA thesis).

Harrison, P. – Clermont, F. (2012), The Influence of LPC Order on the Accuracy of Formant Measurements Across Speakers. In: *Proceedings of IAFPA 2012*, Santander.

Hillenbrand, J. – Getty, L. A. – Clark, M. J. – Wheeler, K. (1995), Acoustic characteristics of American English vowels, *Journal of the Acoustical Society of America*, *97*, pp. 3099–3111.

Labov, W. (1963), The Social Motivation of a Sound Change, *Word*, *19*, pp. 273–309.

McDougall, K. (2006), Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies, *International Journal of Speech, Language and the Law*, *13*, pp. 89–126.

Moos, A. (2012), Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech, *The Phonetician*, *101/102*, pp. 7–25.

Moulines, E. – Charpentier, F. (1990), Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Communication*, *9*, pp. 453–467.

Oppenheim, A. V. – Schafer, R. W. – Buck, J. R. (1999), *Discrete-Time Signal Processing*, Upper Saddle River, NJ: Prentice Hall.

Shue, Y. (2013), *VoiceSauce: A program for voice analysis* (V1.14), Retrieved on October 7, 2013 from http://www.seas.ucla.edu/spapl/voicesauce/.

Sjölander, K. (2014), *Snack Sound Toolkit*, Stockholm: KTH Royal Institute of Technology, Retrieved from http://www.speech.kth.se/snack.

Sjölander, K. – Beskow, J. (2005), *WaveSurfer*, Stockholm: KTH Royal Institute of Technology, Retrieved from http://www.speech.kth.se/wavesurfer/index.html.

Skarnitzl, R. – Volín, J. (2012), Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny, *Akustické listy*, *18*, pp. 7–11.

Viterbi, A. J. (1967), Error bounds for convolutional codes a an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory*, *13*, pp. 260–269.