Part II Reflections

Part II presents reflections that, according to the author's reading of the literature, oppose the understanding of the theory, that is, its intellectual re-enactment and validation.
3 Vowels and Number of Formants

3.1 Inconstant Number of Vowel-Specific Relative Spectral Energy Maxima in Sounds of Back Vowels and of /a–α/

As reported in the literature, when analysing samples of sounds of back vowels and of /a–α/, some sounds may exhibit only one distinct vowel-specific spectral envelope peak, whereas other sounds of the same vowels exhibit the expected two pronounced peaks.

Empirically, the number of vowel-specific relative spectral energy maxima proves to be inconstant for sounds of single vowels.

3.2 Inconstant Correspondence between Vowel-Specific Relative Spectral Energy Maxima and Calculated Vowel-Specific Formant Patterns

If sounds of back vowels and of /a–α/ exhibit only a single vowel-specific spectral envelope peak, according to the literature, formant analysis (e.g. using LPC analysis) often reveals two close formant frequencies. Such cases are therefore referred to as formant merging. It follows that, for the sounds in question, the spectral envelope peak and the calculated first two formants do not correspond to one another.

Yet, if sounds of back vowels and of /a–α/ exhibit two vowel-specific spectral envelope peaks, such a correspondence is generally found.

Thus, the observation of an inconstant number of vowel-specific spectral envelope peaks of sounds of one and the same vowel calls into question the fundamental relationship between spectral envelopes and calculated formants.

No direct parallelism exists between relative spectral energy maxima and calculated formants.

Consequently, formants prove to be constructs of a specific method of analysis (see Section 6.1).
3.3 Inconstant Number of Vowel-Specific Relative Spectral Energy Maxima and of Calculated Vowel-Specific Formants

As shown in Part I, with regard to high front vowels and r-coloured front vowels of some languages, sounds belonging to these vowels can exhibit, in part, similar first and second lower spectral envelope peaks and formant analysis can reveal similar F1–F2. Thus, the sounds of the corresponding vowels are physically distinct only with regard to the third spectral envelope peak and the third formant, respectively.

For such languages, it follows that back vowels, as well as some of the front vowels, are physically describable in terms of different patterns of F1–F2, whereas the remaining front vowels have to be described only in terms of different patterns of F1–F2–F3.

Empirically, the number of vowel-specific relative spectral energy maxima and of calculated vowel-specific formants proves to be inconstant among different vowels.

With regard to spectral envelope peaks, then, the quality of some sounds of back vowels is represented by a single peak, the quality of other sounds of back vowels and sounds of some front vowels by two peaks and the quality of some front vowels by three peaks.

3.4 Addition: “Spurious” Formants

In the spectra of the sounds of certain speakers, an additional spectral envelope peak may occur between the expected first and second or second and third formant. According to the prevailing methodological rules for determining formants, this maximum is not interpreted as vowel specific but as a specific characteristic of the speaker’s voice in question. Therefore, it is referred to as a “spurious” formant.

Such “spurious” spectral envelope peaks also need to be considered within the context of the inconstant number of vowel-specific spectral envelope peaks.

3.5 Addition: “Flat” Vowel Spectra

In the literature, some indications for possible vowel perception related to “flat” spectral parts, lacking any clearly distinctive relative energy maxima, are also given.
3.6 Addition: Inconstant Number of Vowel-Specific Formants in Synthesis

Synthetically produced—and easily recognisable—vowel sounds can be generated for most vowel qualities using three- and two-formant synthesis. For certain vowels, in particular for back vowels and /a–ɑ/, this is also possible by way of a one-formant synthesis.

With regard to synthesised sounds perceived as belonging to one vowel quality, a comparison of the sounds with F1’–F2’ (two-formant synthesis) and the sounds with F1’–F2’–F3’ (three-formant synthesis) reveals differences for F2’, in particular for sounds of front vowels. Similarly, a comparison of the sounds with F1’ (one-formant synthesis) and the sounds with F1’–F2’ (two-formant synthesis) reveals differences for F1’. (However, in the corresponding comparative studies, the fundamental frequency used in synthesis the was not varied systematically.)

Synthesis thus confirms the inconstant number of observable vowel-specific formants. Further, synthesis involving different numbers of formants (different numbers of filters) indicates differences for F1’ or F2’, respectively, although the sounds in question are perceived as belonging to the same vowel.
4 Vowels and Fundamental Frequency

4.1 Fundamental Frequency, First Formant and “Grade” of Vowels

According to prevailing theory, vowel-specific formant patterns are independent of the fundamental frequency of their respective individual sounds.

In general, the frequencies of the first formant of all vowels, as specified in current formant statistics for sounds produced in citation-form words, comparable to relaxed speech, lie within the range of the possible fundamental frequencies for the speakers of a given speaker group. Concerning long German vowels, the lowest statistical values for F1 are given for /i, y, u/, medium values for /e, ø, o/, followed by values for /ɛ, ɔ/ and the highest values are indicated for /a–ɑ/.

If the fundamental frequency involved in producing vowel sounds exceeds the frequencies of the first formant of /i, y, u/ and approaches the frequencies of the first formant of /e, ø, o/, then it is to be expected that the vowels /i, y, u/ become unintelligible because their first vowel-specific formant is no longer physically representable. Thus, the vowels /i, y, u/ would be of a “lower grade”, that is, more restricted in their production, physical representation and intelligibility than the other vowels. The same would apply to /e, ø, o/ compared to /ɛ, a, ɑ, ɔ/ and to /ɛ, ɔ/ compared to /a–ɑ/.

In line with prevailing theory, the possibility that the fundamental frequency of a vowel sound can exceed the first formant frequency of a vowel quality as given in formant statistics leads to the assumption that the “grade” of vowels differs because of vowel-specific acoustic characteristics.

However, everyday experience refutes such a generalising conclusion. If speakers of a given speaker group produce vowel sounds, and if the fundamental frequency of these sounds exceeds the frequencies of the statistically given first formant of /i, y, u/ and approaches the frequencies of the first formant of /e, ø, o/, then all of the six vowels mentioned can be produced with the same “grade” of vowel perception, given speakers with correspondingly good vocal abilities. There is no general impairment of vowel perception for the sounds of /i, y, u/ if the fundamental frequency exceeds statistical F1.
The same holds true—although it is less obvious in everyday utterances and only for good voices—for the vowels /e, ø, o/ produced at fundamental frequencies higher than the statistical values of their first formant frequencies.

Speakers with excellent vocal abilities can even produce clearly intelligible cardinal vowels up to a fundamental frequency that corresponds to the highest statistical F1 of all vowels of the language they master.

In this context, special attention needs to be given to everyday speaking styles or habits that exhibit a fundamental frequency variation of one octave or more. Such styles and habits plainly reveal the significance of the problem of fundamental frequencies above statistical first-formant frequencies, confronting the prevailing acoustic theory of the vowel.

Special attention also needs to be given to utterances of stage voices (in musical and straight theatre, entertainment, film, television etc.) because extensive fundamental frequency variation is one of the hallmarks of the singing and speaking voice in the context of art and entertainment.

Generally, with regard to a fundamental frequency range up to the maximum frequency of the first formant as given in formant statistics, no principally different “grades” of vowel perception in relation to fundamental and first formant frequency can be experienced.

### 4.2 Fundamental Frequency, Spectral Envelope, Formant Pattern and “Grade” of Vowels

If the fundamental frequency of a sound increases, so too does the frequency spacing between the harmonics in the spectrum. As a consequence, determining the spectral envelopes and their maxima becomes difficult. The same applies to the calculation of formant frequencies. According to prevailing theory, it is to be expected that the “grade” of vowel perception is in general also dependent on the fundamental frequency of the sounds: with regard to fundamental frequency, the expected tendency for vowel perception is: the lower, the better; the higher, the worse.

Indeed, considering vowel sounds at higher pitches, many scholars interpret these sounds as related to a spectral undersampling of the formants.

However, one does not only have to consider a general interrelation between fundamental frequency, harmonic spectrum, spectral enve-
lope and expected formant frequencies, but also a formant-specific role within this interrelation: depending upon given statistical frequency values of vowel-specific formants, comparisons show that sounds at higher fundamental frequencies may in some cases exhibit frequencies and relative amplitude maxima of harmonics that correspond to the statistical formant frequencies for the vowels in question, whereas the frequencies of the harmonics of sounds at lower fundamental frequencies lie in between these formant frequencies. For the latter, the formants are subsequently expected to appear as envelope peaks either only indistinctly or not at all, and the corresponding vowel perception is expected to be impaired when compared to sounds at higher fundamental frequencies for which the frequencies of the harmonics match statistical vowel-specific formant frequencies.

Such reasoning leads to the assumption that there is not only a general but also a discontinuous relationship between the intelligibility of vowel sounds and their fundamental frequency: accordingly, vowel sounds at lower fundamental frequencies would, as a rule, be more intelligible than vowel sounds at higher frequencies, but vowel intelligibility would also depend upon the respective relationships between fundamental frequency, harmonic spectrum and vowel-specific formant patterns (as given in formant statistics).

In line with prevailing theory, the relationship between fundamental frequency, harmonic spectrum, spectral envelope and expected vowel-specific formant pattern leads to the same assumption that the “grade” of vowels differs in relation to vowel-specific acoustic characteristics.

However, as explained, everyday experience refutes such a generalised conclusion. Thus, a theory of vowels as elements of language that formulates an inherently qualitative and at the same time discontinuous relationship between fundamental frequency and vowel perception stands in contrast with the—possibly “sensational”—characteristic of a voiced element of language being independent of pitch within the range of intelligible speech.
5 Formant Patterns and Speaker Groups

5.1 Fundamental Frequency, Spectral Envelope, Formant Pattern and “Grade” of Vowels Uttered by Children, Women and Men

If one further extends the reasoning developed in the previous chapter, namely that—according to prevailing theory—the intelligibility of a vowel sound is expected to relate to the respective fundamental frequency of the sound and the (statistically given) first formant frequency of the vowel, then, correspondingly, the “grade” of vowel perception should also depend upon the speaker group: vowel intelligibility should prove to be best for men, average for women and worst for children.

According to prevailing theory, the above relationship between fundamental and first formant frequencies, spectral characteristics and expected differences in the “grade” of intelligibility of different vowel qualities leads to the assumption that the “grade” of vowels varies for different speaker groups (children, women, or men).

Everyday experience also refutes this generalisation. Thus, again, a theory of vowels as elements of language that formulates a inherently qualitative relationship between age and gender on the one hand, and vowel perception on the other, stands in contrast with the—possibly (yet again!) “sensational”—characteristic of a voiced element of language being quasi-independent of a speaker’s constitution (if not impaired).

Vowels as such are related neither to age nor to gender. If direct comparisons of utterances of single speakers show that some speakers produce vowel sounds “better” (better in vowel intelligibility) than others, then, this has to do with the vocal abilities of the individual speakers investigated, not with vowels, speaker groups, or vocal-tract sizes (with the exception of very young children acquiring their first language). As a rule, vowels, as speech sounds of a given language, can potentially be produced with equal intelligibility by speakers of all general speaker groups. Vowels are not attributes of an individual, but elements of language. Vowels are “abstracted” from the individual.
5.2 One Vowel, Different Formant Patterns

In the literature, empirical reference values for vowel-specific formant patterns are given separately for each speaker group (children, women, or men), that is, in group-specific terms (see, for example, Chapter 2). In the first instance, these differences in formant patterns are not explained in terms of varying average fundamental frequencies, but in terms of varying average vocal-tract size.

This view leads to the assumption that each vowel is physically represented by three different speaker group-specific formant patterns, not only in terms of the different fundamental frequencies, but also in terms of the same fundamental frequency: in general, women and men are able to produce clearly recognisable vowel sounds at a child’s fundamental frequency—for instance, at around 250 Hz (see Section 2.1; note, in this context, that in the statistics of Hillenbrand et al., F0 differences between women and children do not exceed 20 Hz). Given such cases of sounds at similar fundamental frequencies, three sounds of the same vowel, produced by a man, a woman and a child respectively, are expected to exhibit three substantially different formant patterns, despite the similarity in vowel perception.

According to prevailing theory, the relationship between vowel-specific formant patterns and age- and gender-related speaker groups leads to the assumption that the physical representation of a vowel is based upon different formant patterns.

Such reasoning also leads to the assumption that women and men are capable of producing sounds of a given vowel with fundamental frequencies substantially higher than those of children, albeit with substantially lower corresponding formant patterns.

The problem that the particular sound configurations in question pose to the theoretical approach discussed here becomes particularly evident when considering corresponding sounds of the vowels /a, ɑ, ɔ, o, u/, which are low-pass filtered with a cut-off frequency of 2 kHz (note that, for these vowels, statistical values of vowel-specific formant patterns F1–F2 for all three speaker groups discussed here are given as ≤2 kHz): then, neither different fundamental frequencies nor different higher spectral energy configurations can play a role in vowel perception and can explain why three different patterns of F1–F2 can be expected to represent the same vowel.
It goes without saying that the above also holds true for the restricted comparison between women and men.

The problem described here becomes particularly acute if, instead of natural vocalisations, corresponding sound configurations are studied by means of vowel synthesis, applying similar fundamental frequencies but different patterns \( F1'–F2' \).

However, in its turn, such a conclusion runs counter the requirement of a psychophysical parallel between perceived vowel quality and physical representation: formant patterns are either vowel specific, which means that clearly distinct formant patterns do not represent the same vowel—regardless of the fundamental frequency—or they are, as such, not directly vowel specific. According to the first stance, the assumption of speaker group-specific formant patterns would have to be questioned. According to the second stance, the assumption of vowel-specific formant patterns in general would have to be questioned.

5.3 Different Vowels, One Formant Pattern

Disregarding the comment in the previous paragraph, the pursuit of the reasoning developed in Section 5.2 leads to the further assumption that a single formant pattern can represent two different vowels: given that the sounds of a vowel produced by a speaker of one speaker group exhibit higher vowel-specific formant frequencies than the sounds of the same vowel produced by a speaker of another speaker group, and that the fundamental frequency plays no substantial role in the physical representation of the vowel in terms of formant patterns, and also given that the vowel-specific formant frequencies of the sounds of the first speaker lie within the frequency range of the possible vowel-specific formant frequencies of the second speaker, then it must be possible to find cases of comparisons of two sounds, each produced by one of these two speakers, that exhibit similar vowel-specific formant patterns, yet are perceived as different vowels.

According to prevailing theory, the relationship between vowel-specific formant patterns and age- and gender-related speaker groups leads to the assumption that a single formant pattern can physically represent two different vowels.

Again, the problem that such sound configurations pose to the theoretical approach discussed here becomes particularly evident when considering corresponding sounds of the vowels /a, ɑ, ɔ, o, u/, because
the vowel-specific formant frequencies of the corresponding sounds of all speaker groups are given in formant statistics ≤ 2 kHz, and in such a frequency range, adults can reproduce sounds exhibiting any of the F1–F2 pattern found in sounds of children. The same holds true when comparing the sounds of men and women.

The problem described here becomes particularly acute again if replicated by means of vowel synthesis, above all including extensive variation of the fundamental frequency.

However, in line with the explanation given above, the assumption of a possibility of twofold representation, according to which a single formant pattern can correspond physically to the sounds of two different vowels, runs counter to the requirement of a psychophysical parallel between perceived vowel quality and physical representation. At the same time, indeed, it directly contradicts prevailing theory.

This consideration engenders a decided scepticism about the claim that vowel-specific formant patterns are both fundamentally and continuously dependent upon the speaker group, that is, upon vocal-tract size. A fundamental dependence is already difficult to understand from an intellectual standpoint because, as mentioned, vowels do not “have” an age or gender. Besides, the simple fact that sounds of back vowels can be synthesised at fundamental frequencies, observable in sounds of children as well as in sounds of men, paradigmatically illustrates the problem: if, in synthesis, F1–F2 is changed substantially but the fundamental frequency is held constant, in general, the perceived vowel quality also changes, irrespective of whether the F1–F2 of the synthesis corresponds to a pattern observed for natural sounds of a child or of a man.

At the same time, the above reflection suggests an alternative explanation for the existing empirical findings, which seemingly provide evidence for speaker group-specific formant patterns: vowel-specific spectral energy configuration, and with this this calculated formant patterns, can depend upon fundamental frequency.

It is remarkable that, in general, formant statistics deemed worthy of reference in the literature do not give frequency values of formant patterns of the different speaker groups for systemically varied fundamental frequencies. Thus, currently, there is no empirical evidence in the literature to support the claim that observed, speaker group-specific formant patterns of vowels should in principle not be attributed to the different—and simultaneously observed—fundamental frequencies of the respective sounds but, instead, to different average vocal-tract
sizes. With regard to the first formant for all vowels, and probably also to the second formant for back vowels, the present reflection indicates that such evidence cannot be furnished.

5.4 A Gap in the Reasoning

As indicated, existing formant statistics suggest that, irrespective of fundamental frequency and perceived vowel quality, adults are capable of producing sounds for almost all variants of F1–F2 patterns as found in children’s vowels. Thus, even though adults have larger vocal tracts than children, for most vowels, they are nevertheless capable of producing sounds that exhibit the same vowel-specific formant patterns, above all F1–F2, as evidenced for the sounds of children.

If it is indeed the case that speakers of all three speaker groups are considered to be capable of producing the same vowel-specific patterns for a substantial part of vowels, then how are the pattern differences discussed above to be understood? (Many scholars assume that the schwa sound defines the midpoint of a speaker’s vowel space and plays a central role for the formant pattern differences discussed: because of different average vocal tract lengths and different resonance patterns of related open tubes of speakers of different age and gender, it is deduced that different vowel-related format patterns mirror different midpoint reference patterns. However, in the present context, such an assumption does not dispense from the question posed: sounds of schwa, too, can be produced on different fundamental frequencies, and the independence or dependence of related formant patterns on fundamental frequency for perceptually unaltered schwa quality has not yet been clarified.)

Even though existing statistical values list vowel-specific formant patterns for children exceeding those for adults, and for women exceeding those for men, there are nevertheless exceptions: in some cases, as shown by some statistics, single vowel-specific formant frequencies, or even vowel-specific formant patterns F1–F2 or F1–F2–F3, for sounds produced by men do not differ from those for sounds produced by women; they may even slightly exceed the latter. (Thus, remarkably, the formant patterns given by Fant, 1959, for a single male and a single female speaker do not show a consistent speaker group related difference; see Section 2.1, Table 3. Besides, there are cases in which the statistical F1 of women slightly exceeds the F1 of children, see, for instance, Section 2.1, Table 2, and the corresponding values for the vowel /ʌ/.) This raises the same question as above.
The relationship between vowel-specific formant patterns and age- and gender-related speaker groups described in terms of prevailing theory fails to explain why, despite different vocal-tract sizes, similar vowel-specific formant patterns are basically possible at least for the majority of vowels but are—according to theory—not realised (actually not produced).

In addition, this formulation could also prove to be generally applicable: it could prove to be the case that all vowel-specific formant patterns, F1–F2 and F1–F2–F3 as given in formant statistics for children, can also be produced by women and men. (With regard to this aspect, utterances of voice-over artists are of particular interest.)

Repeating and insisting: given a psychophysical perspective, the correspondence between intelligible vowel sounds and the vowel-related physical characteristics must be formulated as such. The formulation of speaker-independent and, in a strict and direct sense, vowel-specific acoustic features represents the touchstone for any acoustic theory of the vowel.

5.5 Addition: Formant Patterns of Voiced and Whispered Vowel Sounds

Empirical studies comparing voiced and whispered vowel sounds indicate substantial differences in the formant patterns related to the perceived vowel qualities. In particular, the first formant frequency of whispered sounds of a given vowel (and, according to some studies, the second formant frequency, too) are found on significantly higher frequency levels than those of voiced sounds. (As mentioned in Section 1.4, such differences are explained as a consequence of differences in the geometry, and thus the resonances, of the glottal area of the vocal tract for the two different phonation types in question.)

This finding relativises again the attempt to establish a direct correspondence between vowels and formant patterns: the sounds of the same vowel can exhibit different formant patterns, not only because of different average vocal-tract sizes but also because of different kinds of phonation acting upon a configuration of a single vocal tract.

Moreover, comparisons between published formant frequencies of whispered and voiced vowel sounds indicate that all F1, and the majority of F2 ≤ 1.5 kHz, of whispered sounds produced by men generally exceed the corresponding F1 and F2 of voiced sounds produced by women, given the same perceived respective vowel identities and notwithstand-
ing men's larger vocal tract. The same applies to a comparison between whispered sounds of women and voiced sounds of children. Restricted to F1, this also applies to the comparison between whispered sounds of men and voiced sounds of children.

This observation relativises in turn the assumption of a correspondence between vocal-tract size and vowel-specific formant patterns: based on the values given in the literature, such a correspondence is documented only for sounds of one and the same phonation type, not for a comparison of sounds of different phonation types. Besides, it should be noted that the frequency differences of the lower formants for the sounds of a given vowel, which relate to different types of phonation, e.g. voiced versus whispered sounds, are in general greater than the corresponding formant frequency differences between the different speaker groups.

Thus, most importantly, vowel-related formant patterns produced by one vocal tract can differ more than vowel-related formant patterns produced by different vocal tracts with very different tract sizes.

Moreover, referring to Section 5.3, a single formant pattern seems able to physically represent different vowels not only if the corresponding sounds are produced by speakers belonging to different speaker groups, but also if an individual speaker varies his or her phonation.

Such consideration will be discussed further in Part III: comparisons between the formant patterns of voiced and whispered sounds, as documented in the literature, refer only to the average (lower) fundamental frequency of voiced vowel sounds produced in citation-form words, but not to a comparison including a systematic variation in fundamental frequency of voiced sounds. (Such an experimental arrangement assumes, once again, that formant patterns are independent of fundamental frequency and are, therefore, negligible when comparing voiced and whispered sounds.)
6 Terms of Reference, Methods of Formant Estimation

6.1 Formant and Sound Spectrum

Given that the terms “resonance” and “formant” are distinguished from each other, as a means of distinguishing the characteristics of the vocal tract from those of the sound spectrum, then the psychophysical question of the vowel relates to formants only. According to prevailing theory, it is assumed that, in the first instance, the spectrum of a vowel sound exhibits determinable relative energy maxima, which are related to vowel-specific frequency ranges, and that, as a rule, the frequencies of these relative spectral energy maxima correspond to calculated formant frequencies, for example, applying LPC analysis. (Note that, nowadays, formant frequencies are no longer derived as numerical values from the spectral envelope but, instead, are calculated as filters of an analytical model, although the corresponding numerical results are in many cases crosschecked on the basis of a spectrogram.)

As discussed in Sections 3.1 and 3.2, the sound spectra of back vowels and of /a–ɑ/ can exhibit only one single vowel-specific spectral energy maximum, although formant analysis using an analytical model (e.g. LPC analysis)—under involvement of “phonetic knowledge” and sometimes with interactive manual adjustment of parameter settings—indicates two vowel specific formants, often close in frequency. This contradicts the assumption that the number and frequency of relative spectral energy maxima, that is the envelope peaks, always correspond to analytically determined formants.

As mentioned in Section 4.2, due to the increasing frequency spacing of the harmonics, the higher the fundamental frequency, the more difficult it becomes to determine the spectral envelope and its peaks (for further details, see also Section 6.4). This in turn impedes the formulation of a general correspondence between relative spectral energy maxima and calculated formant frequencies.

Regarding the current procedures used in formant analysis and the corresponding numerical values of formant patterns, it follows that in many cases—and thus in principle—the term formant often does not designate a characteristic of the sound spectrum itself, but instead a construct or even artefact of the respective method of analysis.
In the current literature, the term formant—if distinguished from resonance—generally refers neither to any actual characteristic of the vocal tract nor to any actual characteristic of the sound spectrum. The term generally refers to filters of an analytical model. At the same time, formants are not determined on the basis of spectra but on the basis of such an analytical model.

Thus, the assumption that a direct correspondence exists between resonances as a physical property of the vocal tract, spectral energy maxima as a physical characteristic of the vowel sound produced and filter frequencies derived from methods used in the acoustic analysis of vocal sounds, loses its plausibility.

### 6.2 Speaker Group and Vocal-Tract Size

As discussed, prevailing theory supposes a relationship between vowel-specific formant patterns and age- and gender-related speaker groups and explains corresponding differences in terms of the respective average vocal-tract sizes.

It can be assumed that some women have larger vocal tracts than some men. Comparing the vowel sounds of these female and male speakers, the following constellation is of particular interest in the present context: the sounds of the female speakers in question exhibit fundamental frequencies corresponding to the average fundamental frequency values for women in general, as given in formant statistics, and the sounds of the male speakers in question exhibit substantially lower fundamental frequencies. Then, according to prevailing theory, the vowel-specific formants of these female voices would have to exhibit lower frequencies—despite comparatively higher fundamental frequencies—than the corresponding formant patterns of these male voices.

Extending such consideration, this comparison raises the question of a systematic investigation of the relationship between vocal-tract size and vowel-specific formant patterns within a single speaker group.

Besides the lack of an empirical basis for the questions raised here, the above reflections again point to the fact that prevailing theory does not claim that vowel-specific formant patterns depend in principle on age and gender, but that different vowel-specific formant patterns exist for different vocal-tract sizes: prevailing theory only refers to speaker group-specific differences in average vocal-tract sizes.)
The term “age- and gender-related speaker group” is related to the term “age- and gender-related average vocal-tract size”.

### 6.3 Formant Analysis and Objectivisation

Concerning natural vocalisations, current analytical methods for determining formants apply a model-like procedure in order to calculate a specific configuration of source sound and filters which, by means of transformation of source by filters, “reproduces” a sound that best corresponds to the real sound. (The same applies to whispered vowel sounds, in relation to the source as noise.)

Such a procedure must not only assume certain characteristics of the source sound but also a certain number and certain characteristics of the filters involved in the frequency range under investigation. (Note that, according to prevailing theory, different numbers of formants are expected for a given frequency range in relation to different speaker groups because of their different average vocal-tract size. Thus, the number of filters for the analysis of a sound must be set accordingly.) How closely the characteristics of the source sound approach actual phonation remains open. The same applies to the question of whether the number of filters and their characteristics actually correspond to real articulation and its resonance.

Thus, formants cannot be determined reliably on the basis of a vowel sound alone. Analysis requires at least some prior knowledge of whether the sound under investigation has been produced by a man, woman, or child, assuming that this information is sufficient to deduce the number of filters (related to the frequency range of interest) to be used in formant analysis.

Besides, subsequent automatically calculated formant frequency values are often double-checked visually on the basis of the sound spectrogram: if the values calculated in the first step—based on analytical parameters according to existing standards and known speaker group—do not correspond to the relative spectral energy maxima of the analysed sound, then the number of filters is varied and analysis is performed until such a correspondence occurs. As a rule, the characteristic of the source sound is not altered. However, this only applies to cases where such an interactive analysis is able to produce vowel-specific numbers and frequencies of formants that correspond to the number and frequency ranges to be expected according to prevailing theory and established statistical patterns, and which are also clearly indicated in the spectrogram. If an interactive procedure of ana-
ysis yields no values with such a correspondence, then the respective vowel sounds are often excluded from further studies, irrespective of vowel perception. Exceptions include so-called “formant merging”, as discussed in Section 3.2.

Thus, current methods of formant analysis presuppose that researchers have the necessary analytical skills, that is, a knowledge of the existing phonetic principles and rules of interpretation as well as extensive first-hand experience of conducting such an analysis. This involves prior training because such an analysis involves contextual knowledge, the ability to visually compare numerical values with a corresponding sound spectrogram, together with the ability to interpret the latter visually, and also the skills to vary filter settings interactively and to perform the repetition of numerical analysis. Consequently, methods of formant analysis are not completely objectifiable. If they were, then researchers would play no part as individuals in such research.

Strictly speaking, methods of formant analysis are not fully objectifiable; accordingly, they cannot be fully automated.

Most importantly, these procedures are also very time consuming. Therefore, investigations based on very extensive samples of sounds are problematic with regard to method. This is the case particularly if the fundamental frequency is varied: then, specific problems of analysis aggravate the costly character of the method as such. Obviously, this holds true for all repetitions and verifications of existing investigations.

6.4 Formant Analysis, Fundamental Frequency and Speaker Group or Vocal-Tract Size

In addition to formant analysis not being fully objective and automated, it also depends on the respective fundamental frequencies of the sounds. To repeat: the higher the fundamental frequency, the more difficult it becomes to determine the spectral envelope peaks expected because the frequency spacing between the harmonics become too large to accurately define the spectral envelope. It also becomes increasingly difficult to determine the formants within any of the existing analytical frameworks.

With regard to critical limits of fundamental frequencies, above which methods of formant analysis become unreliable, two kinds of reference
values need to be considered: firstly, half the frequency of the lowest first formant for a speaker group in terms of an average vocal-tract size, and secondly, the frequency of the lowest formant for a speaker group.

For a fundamental frequency above half of the first formant frequency (F0 > ½F1), the frequency spacing between the harmonics is already so extended that defining a spectral envelope and evaluating the calculated numerical formant frequencies becomes problematic. (Note that for such sounds, the formants may not be clearly indicated by at least two harmonics.) According to this first kind of limit, and referring to the standard values established by Hillenbrand et al. (1995) for F1 of /i/ (the lowest average value for F1 in these reference statistics), formant analysis becomes critical for fundamental frequencies higher than:

- 226 Hz for sounds of children (involving short vocal tracts)
- 219 Hz for sounds of women (involving medium-sized vocal tracts)
- 171 Hz for sounds of men (involving long vocal tracts)

For a fundamental frequency above the lowest first (statistically given) formant frequency for a given speaker group, under the assumption of independence of formants from fundamental frequency, it is basically impossible to distinguish all F1 of all vowels produced by speakers of that group, not to mention the aggravated problem of determining the spectral envelope. According to this second kind of limit, and again referring to the above statistics, methods of formant analysis lack a methodological basis for fundamental frequencies higher than:

- 452 Hz for sounds of children (involving short vocal tracts)
- 437 Hz for sounds of women (involving medium-sized vocal tracts)
- 342 Hz for sounds of men (involving long vocal tracts)

Note that referring to the statistics of Pätzold and Simpson (1997) for German vowels, shown in Section 2.2, the limits would have to be set even on lower frequencies: ½F1 of /i/ corresponds to 165 Hz for women (medium-sized vocal tracts) and to 145 Hz for men (long vocal tracts), respectively; F1 of /i/ corresponds to 329 Hz for women and to 290 Hz for men or long vocal tracts, respectively.

In this context, attention should also be given to the fact that, according to several formant statistics, the frequency distance between F1 and F2 for sounds of some back vowels is given <500 Hz. Thus, the frequency spacing of the first two harmonics in a spectrum of a sound
on a fundamental frequency above this frequency limit exceeds the F1–F2 distance mentioned, which renders formant estimation obsolete within the existing theoretical framework.

The first lists of frequency limits given above for $F_0 > \frac{1}{2}F_1$ suggests that methodologically speaking the analysis of vowel sounds of children and women must be considered problematic in general. The critical fundamental frequency value mentioned for children is considerably lower than the empirically determined average fundamental frequency that children exhibit when producing vowels in citation-form words, which can be considered as related to relaxed speech on a comparatively low fundamental frequency (see, for example, the statistics in Section 2.1). Thus, most vowel sounds produced by children in their everyday expression, exhibit substantially higher fundamental frequencies.—According to Hillenbrand et al. (1995), the mentioned critical fundamental frequency value for women corresponds to the average fundamental frequency of women producing vowels in citation-form words. In everyday speech, however, vowel sounds in a fundamental frequency range of up to one octave higher than this value are the norm. Moreover, according to Pätzold and Simpson (1997), the mentioned critical fundamental frequency value for women is again considerably lower than the average fundamental frequency generally given in vowel statistics.—The problem discussed here seems to be less pronounced among men than among women and children, but it nevertheless concerns a substantial part of their utterances.

The second list of frequency limits reveals that, for methodological reasons, any determination of formant patterns of vowel sounds exhibiting fundamental frequencies that exceed low first-formant frequencies does not make sense, since general rules for formant estimation can no longer be formulated. In this regard, particular consideration needs to be given to voices exhibiting extensive prosodic variations in fundamental frequency, which can be experienced in everyday speech and, very pronounced, in the field of art and entertainment. (Noticeable, with regard to everyday speech, the literature does not provide ample documentation of the occurrence and significance of such extensive variation in fundamental frequency, allowing for a validation of the significance of the methodological problem of formant estimation discussed here. However, in the Materials section, examples of corresponding utterances are documented; see Section M8.2.)
Within the prevailing theoretical framework, the reliability of formant analysis depends on fundamental frequency and the age- and gender-related speaker group, that is, vocal-tract size. Depending on the latter, formant frequency estimation becomes critical for fundamental frequencies above c. 175 Hz, and formant frequency estimation can no longer be methodologically substantiated for fundamental frequencies substantially above 350 Hz. Consequently, formant analysis cannot be applied to all cases of clearly intelligible vowel sounds.

A part of the literature tends to equate the methodological problem with a particular characteristic of vowel perception, which leads us back to the two assumptions discussed in Sections 4.1 and 5.1: firstly, that vowels produced by children and women are basically less intelligible than those produced by men; and secondly, that at least some vowels of sounds at a fundamental frequency substantially above 350 Hz can no longer be clearly distinguished. As suggested, however, both assumptions contradict actual vowel perception.

6.5 Addition: Parameter Adjustments in Formant Analysis and Inconsistent References to Vocal-Tract Size

On the one hand, formant parameters in current procedures of formant analysis are defined prior to analysis of the sounds depending on the corresponding speaker group, that is, the assumed average vocal-tract size of the speakers. On the other hand, these parameter settings are sometimes interactively altered during the procedure if the calculated numerical values do not yield the expected number of formants in the expected vowel-specific frequency ranges compared to the respective spectrogram.

Thus, for example, with regard to sounds of a single speaker, LPC analysis involving standard parameters according to the related speaker group (average vocal-tract size) may yield the expected values for only a part of the sounds, whereas the analysis of other sounds may require the parameters to be set to the standard of another speaker group (average vocal-tract size) or to a setting that is entirely different from any speaker-group related standard given in the literature.

This reveals an inconsistency in how parameter settings are established: in the first instance, default settings of analytical parameters are related to specific vocal-tract sizes, whereas any corrections of these settings are related to the respective general (not vocal tract related) degree of “formant resolution” of the analysis.
6.6 Addition: Spectrum, Formant Pattern, Resynthesis

As explained in Section 6.1, current methods of analysis yield no consistent and direct relationship between spectrum, spectral envelope and formant frequencies. Consequently, this raises the question of the existence of a general relationship between a natural vowel sound, the determined formant pattern and resynthesis.

Currently, resynthesis is indeed being used to examine the reliability of calculated formant patterns. However, this kind of verification is unable to substantially relativise the general problems of the existing methods of analysis described above: resynthesis is feasible only if formant analysis is not fundamentally at issue and only with regard to a limited variation of analytical parameters.

Moreover, the question of resynthesis must be discussed against the background of synthesised sounds as discussed in Section 3.1, indicating the possibility of substantial differences in formant patterns of sounds of one vowel: if a certain analytically determined formant pattern used in a resynthesis reveals an “expected” vowel identity in a perceptual test, then this does not mean that another determined formant pattern, based on a different parameter setting, and applied in a second resynthesis, in principle cannot reveal the same vowel identity. Further, the possibility cannot be excluded that there are cases of sounds for which, with regard to the perceived vowel quality, based on “unexpected” formant patterns may produce a better approximation to the quality of the natural sounds in question than based on “expected” formant patterns.

6.7 Addition: Formant Analysis and Objectivity with Regard to Synthesised Vowel Sounds

It is noteworthy that, if a sound is synthesised using a specific pattern of filters and filter bandwidths, the formant pattern of a subsequent analysis may differ from the synthesis filters if the number of filters used is not communicated to the scholar conducting the analysis.

Moreover, the problem of possible differences of filters used in synthesis and formant patterns obtained in analysis will be substantially enhanced if the fundamental frequency is varied independent of the filters.
6.8 Addition: Formant Patterns and Resynthesis outside of the Framework of Prevailing Theory

It is also noteworthy that, if formant patterns are calculated outside the framework of prevailing theory, for example, using LPC analysis as a method to decompose any sound into a source and a set of filters, irrespective of the fundamental frequency and the perceptual quality and not relating the decomposition to existing formant or resonance statistics (and therefore not considering a direct relationship between spectral peaks and resonances of the vocal tract), and if the results of analysis are used in resynthesis, for many examples of natural utterances, resynthesis reproduces similar intelligible vowel qualities, even for very high fundamental frequencies. Obviously, then, formant patterns will sometimes deviate strongly from the statistical patterns given in the literature.