Principles of Phonetic Segmentation

Book · January 2013

CITATIONS
12

READS
6,122

2 authors:

Pavel Machač
Charles University in Prague
27 PUBLICATIONS **198** CITATIONS
SEE PROFILE

Radek Skarnitzl
Charles University in Prague
84 PUBLICATIONS **746** CITATIONS
SEE PROFILE

PRINCIPLES OF PHONETIC SEGMENTATION

Pavel Machač & Radek Skarnitzl

1. Introduction

1.1. Why do we need segment boundaries?

The ultimate goal of any phonetic research is to understand the structure of speech and its various functions in communication (Kohler, 2007). To reveal the structure, we must try to find a sensible and generally acceptable way of delimiting the primitive units of this structure. In practical terms, we need to divide the continuous acoustic signal into discrete segments and associate them with more or less abstract phonetic symbols. Obviously, the size of the units depends on the nature of the research task at hand: we may be interested in segmenting, for example, speechsounds, words, stress groups, intonation phrases, or breath groups.

In this book, we will focus on the segmentation of units on the level of speechsounds. One might argue (and we have encountered this argument) that the knowledge of segment boundaries is not necessary for most areas of phonetic research. It is true that some specific research tasks require other units or parameters. We believe, however, that the knowledge of segment boundaries is still the most universal way to approach the speech material. Annotation on the level of individual segments will be useful not only for studying segmental properties of speech (e.g., temporal characteristics, spectral changes within a speechsound), but also for many kinds of tasks associated with what we call prosodic research. Let us look at only two examples: (1) to examine intonation patterns (not mere F0 contours) we want to know the temporal midpoints of syllable nuclei; (2) the investigation into rhythmic properties of a language is usually related to the temporal behaviour of speechsounds or their classes.

It is well known that one sentence will never be pronounced twice, from the objective physical viewpoint, in an absolutely identical way. Obviously, various speakers will differ in their productions, but even the same speaker in the same communicative and semantic context will not produce two completely identical sentences. In short, speech is an extremely variable phenomenon. The purpose of phonetic investigations is to find some stability, invariance in this variability, because if some degree of invariance did not exist, speech could not function as a means of communication.

Invariance in speech cannot be revealed by examining a few sentences uttered by one speaker. What we need is a representative sample of speech material, a large and structured corpus. To be able to talk about a **phonetic corpus**, the recorded speech must be processed in a uniform way. For our purposes, this processing includes not only transcription, but especially segmentation.

The demarcation of phonetic units - whether segments or others - can proceed in two ways: automatically or manually. A number of automatic instruments have been developed, most frequently based on HMMs (e.g., Wester *et al.*, 2001; Kominek *et al.*, 2003; Pollák *et al.*, 2007).

Unfortunately, these methods are at present not accurate enough for phonetic research and they need manual correction. An HMM-generated segmentation and a manually corrected segmentation of two words are compared in Figure 1.1 (this serves as an illustration, and the discrepancies will not be analyzed here). It is obvious that the output of HMM segmentation can be used for a rough indication of segment boundaries, but not for drawing linguistically interpretable conclusions. This leads to our conviction that human input is essential in the preparation of speech corpora, if we have truly phonetic research in mind. Human input here entails a manual approach to segmentation.

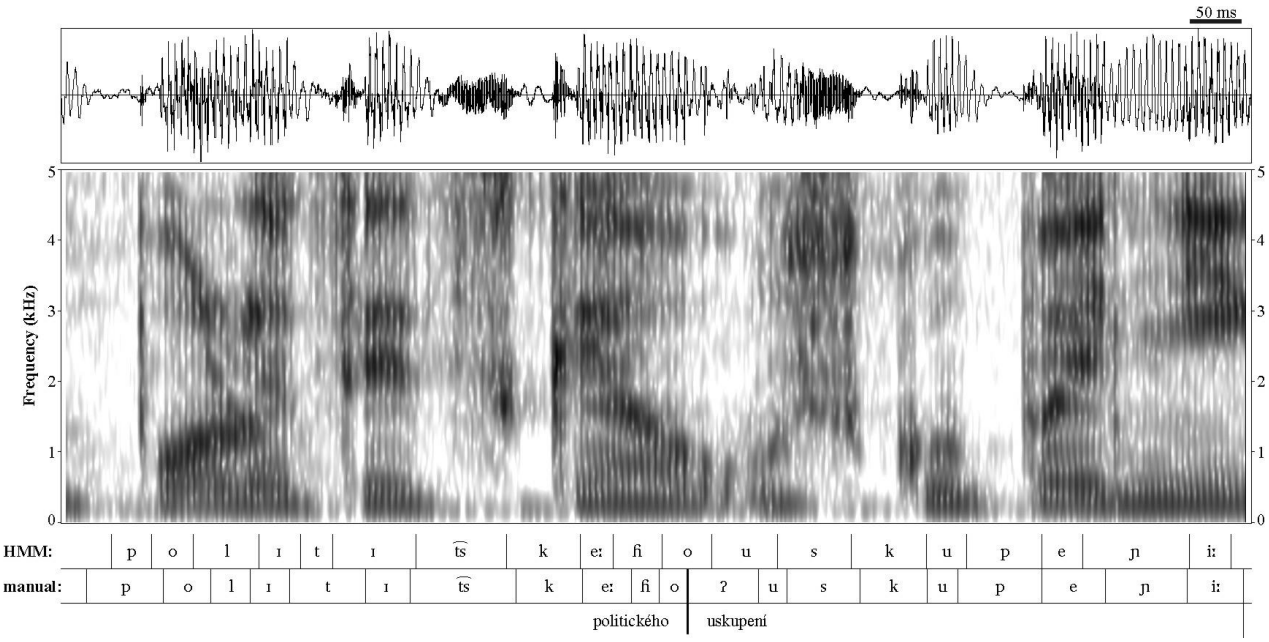


Figure 1.1. Comparison of HMM-generated and manually corrected segmentation of two Czech words.

Naturally, we are aware that manual segmentation has several disadvantages. First, it is known to be time-consuming, and developing a phonetic corpus is thus always a long-term endeavour. Second, manual segmentation is demanding in terms of labeller expertise. Many researchers have criticized it as inherently subjective and therefore inconsistent and irreproducible (e.g., Wesenick & Kipp, 1996; Pitt *et al.*, 2005). Everyone who has attempted to manually segment a stretch of speech has probably had the bitter experience of not being able to decide on the location of a segment boundary. More frequently than we would like, there seem to be several plausible reasons for considerably different boundary placements, or there seem to be no cues for boundary placement at all. Finally, we make a decision and, returning to the same item the following day, change our mind and move the boundary elsewhere. This means that both inter-labeller and intra-labeller consistency is an issue in manual segmentation.

The accuracy of manual segmentation across different labellers has been examined in various studies. Cosi *et al.* (1991, quoted in Pauws *et al.*, 1996) showed that more than 10 % of boundaries differed in their placement by more than 20 ms. The results of inter-labeller comparison in Pitt *et al.* (2005) show an average deviation in boundary placement of 16 ms, and those in Wesenick & Kipp

(1996) a deviation of about 10 ms. Kvale & Foldvik (1991) labelled 748 speechsounds based on relatively simple criteria and found that 96.5 % of boundaries had a deviation of less than 20 ms.

Several years ago, we decided to try to minimize inter-labeller discrepancies. We wanted to see whether relatively simple guidelines for labellers, based on (if possible) phonetically significant events in the acoustic continuum, can lead to a higher inter-labeller agreement. We formulated guidelines for specific speechsound combinations: intervocalic plosives, fricatives and nasals (Volín *et al.*, 2008). Mean deviations across three labellers turned out to be significantly lower than in the comparable study of Wesenick & Kipp (1996), as shown in Table 1.1.

boundary type	***mean deviation (ms) Wesenick & Kipp (1996)***	***mean deviation (ms) Volín et al. (2008)***
vowel-plosive	12.0	1.8
plosive-vowel	6.0	1.3
vowel-fricative	8.0	3.0
fricative-vowel	9.5	2.4
vowel-nasal	9.0	2.0
nasal-vowel	8.0	2.6

Table 1.1. Comparison of mean inter-labeller deviations in Wesenick & Kipp (1996) and in Volín *et al.* (2008). For simplification, the differences between voiced and voiceless obstruents are not listed here.

To be able to compare our results with those of Cosi *et al.* (1991, as reported in Pauws *et al.*, 1996), the deviations in boundary placement are expressed in terms of increasing correct margins in Table 1.2. Although the results of Cosi *et al.* are presumably based on all segment combinations, it is obvious that segmentation guidelines can markedly reduce inter-labeller discrepancies.

correct margin	***intervocalic plosives***	***intervocalic fricatives***	***intervocalic nasals***
= 0 ms	53 %	32 %	43 %
< 3 ms	82 %	66 %	74 %
< 6 ms	96 %	88 %	91 %
< 9 ms	98 %	95 %	96 %
< 15 ms	99.4 %	99 %	98 %

Table 1.2. Correct margins in the segmentation of intervocalic plosives, fricatives, and nasals (based on Volín *et al.*, 2008).

With such encouraging results, we decided to formulate similar segmentation rules for other speechsound combinations and to gather them in the present study. The result of our effort is what you are just about to explore. We believe that the existence of such rules will allow more people (even students) to work on the development of a phonetic corpus, while guaranteeing (at least to a point) a uniform approach to segmentation. This will speed up the preparation of the corpus without

compromising the reliability of segmentation. Our inter-labeller reliability will be addressed in the final section of the book.

Stipulating segmentation guidelines has been attempted before, for example by the creators of the Buckeye corpus who published an online labelling manual (Kiesling *et al.*, 2008). This manual is a set of written instructions, without any illustrations of spectrograms or waveforms, and some of the guidelines are, in our opinion, not sufficiently descriptive. We tried to specify the criteria for boundary placement as rigorously as possible, and to accompany them by visual examples.

1.2. What do we mean by "the boundary"?

While creating this handbook, we have always kept in mind the fact that the exact location of boundaries in the speech signal continuum is often illusory. In the ideal case, the contrast between segments is so salient that boundary location seems to be relatively straightforward. Even untrained labellers then manifest comparatively high degree of agreement, for example in sequences of vowels with voiceless obstruents. It is in situations when the contrast of neighbouring speechsounds is inherently low, and/or when the transition phase between them is gradual, that problems arise. As an example, we can mention the sequence of a vowel and an approximant. In any case, it can be said that careful pronunciation is reflected in the acoustic signal by a higher contrast between segments of any type than in careless, implicit pronunciation. Implicit pronunciation is characterized by an undershoot of the canonical articulatory targets, by a looser synchronization of glottal and supraglottal activities, and of individual articulators in general. Transition stages between speechsounds tend to be relatively slow, which means that there is a massive (bordering on complete) overlap of the phonetic features (see section 1.3) of neighbouring speechsounds.

We shall allow ourselves a small digression and, with some tolerance, imagine the pseudo-three-dimensional display of sound, the spectrogram, as the Earth's surface and segment boundaries as borders between states ("political entities"), in other words boundaries which are artificial yet necessary.

It is obvious at first sight that the location of some state borders is, apart from political reasons, motivated especially by prominent geographical landmarks, a river or a mountain range for example, i.e. some kind of a natural boundary based on a salient contrast on the surface. Sometimes the location of state borders cannot be based on geographical landmarks since the terrain changes are very gradual. However, the political organization of neighbouring, autonomous entities requires borders, and their location is then determined as a result of political negotiations and decisions (or, in a grim scenario, as a result of a violent solution).

When segmenting the speech signal, we also exploit salient boundaries in the spectrogram and waveform; that is, we start with the richness of physical reality. However, as speakers coarticulate (begin pronouncing one speechsound while still finishing the preceding speechsound), we tend to find in the speech signal shorter or longer gradual transition phases, rather than clear-cut boundaries between speechsounds. On the one hand, the existence of individual speechsounds in the signal (the

“political entities”) is obvious, on the other hand we need to employ “political negotiations” to determine rules for boundary location in transition phases and, subsequently, to make an arbitrary decision, at least to a certain extent.

For these reasons, the purpose of this treatise is not to define the exact place in the signal where one speechsound ends and another one begins, but to provide guidelines for situations when there is a requirement on uniform segmentation, such as when more people work on the same corpus which is to be used in phonetic research. Some of the decisions concerning boundary placement may be somewhat arbitrary; that, however, is better than no decisions at all.

If we look at boundaries in a slightly more general way, we may regard them as real or imaginary dividing lines between subsystems within a system. Elements of subsystems manifest certain common features or comparatively stabilized relationships which are at the same time significant for their differentiation from other subsystems.

In certain cases, each of the neighbouring subsystems occupies its own space whose demarcation is relatively easy (Figure 1.2a). One can easily imagine, though, that elements of one subsystem will occur in places where peripheral elements of another subsystem can be found as well (Figure 1.2b). An accurate separation of neighbouring subsystems can thus hardly be unambiguous. Boundary locations are then a result of intellectual activity, and the degree of their substantiation can be disputed.

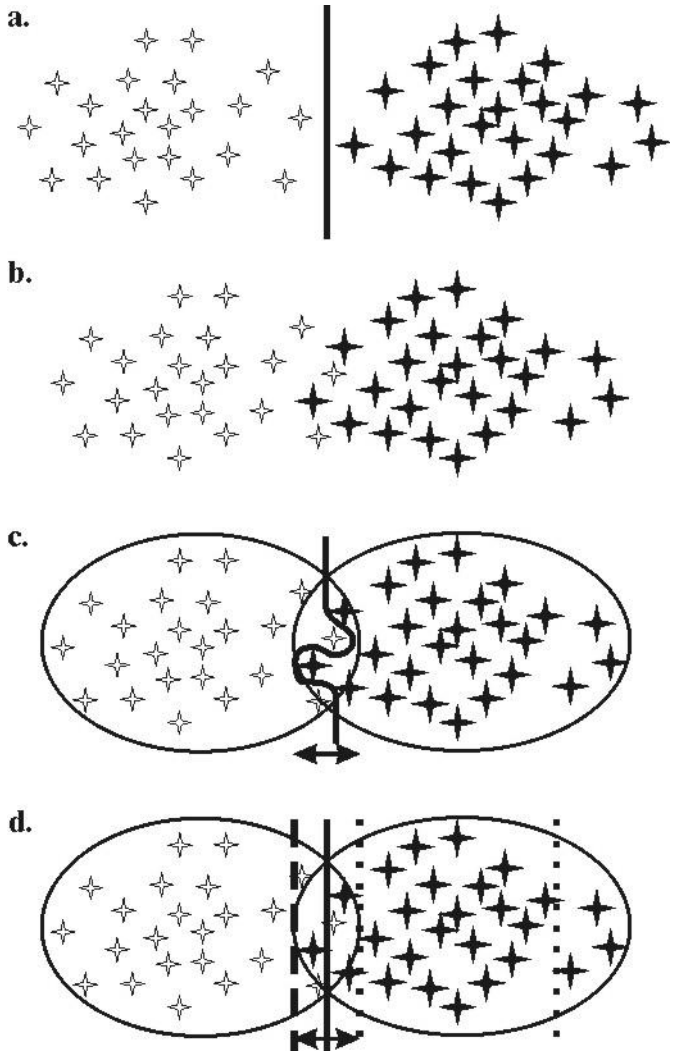


Figure 1.2. Options for boundary location between neighbouring subsystems (see text).

Creating boundaries often helps us to orient ourselves in complex, variable systems. If we draw the boundary exactly between individual elements of the two indiscrete subsystems (Figure 1.2c), we arrive at a result which is complicated to describe and unsuitable for further processing. Frequently, we need to express the boundary between neighbouring subsystems by means of one value of the given physical quantity (for instance, one point on the time axis in the spectrogram), although we are aware of the transition area which includes elements of both subsystems (see the double-sided arrow on the imaginary horizontal axis in Figure 1.2c).

We may imagine a specific utterance as a system of (nearly) linearly concatenated, more or less indiscrete subsystems, or speechsounds. Each pronounced speechsound can, in turn, be regarded as a system of elementary sound impulses whose arrangement characterizes the given speechsound acoustically and makes it identifiable and recognizable perceptually from neighbouring speechsounds.

Let us mention the options when deciding on the boundary location between two speechsounds from the above-mentioned system:

1. Assuming the two subsystems are of equal importance, the boundary will be placed in the middle of the transition area (see Figure 1.2d - solid line).
2. If we consider one subsystem to be dominant (e.g., the filled asterisks), the boundary will be placed outside of this subsystem (see Figure 1.2d - dashed line).
3. If we consider the centre of one subsystem to be dominant (e.g., the filled asterisks), the boundary will be placed outside the centre of that subsystem (see Figure 1.2d - dotted line), while peripheral elements remain beyond the boundary.

When segmenting the speech signal, we make use of all the three options mentioned above, for example between the following speechsound types (see relevant chapters):

1. vowel - approximant;
2. fricative - plosive (the fricative noise is considered to be dominant);
3. vowel - obstruent (the formant structure of the vowel is regarded as the centre, the onset and offset of F0 are regarded as periphery).

As we have mentioned above, our objective is to increase the reliability of manual segmentation by means of stipulating relatively simple and unambiguous rules. Although such rules will, to a certain extent, be arbitrary, we try to base them on inherent phonetic characteristics of the given types of speechsounds.

1.3. Phonetic features

By phonetic features, we mean articulatory, acoustic, and perceptual properties which characterize every speechsound from two viewpoints:

a) the speechsound as a **theoretical construct**,
b) **specific realization of the speechsound** in the utterance.

For segmentation purposes, it seems useful to distinguish between inherent and extrinsic phonetic features, as described in the following subsections.

1.3.1. Inherent phonetic features

The speechsound as a theoretical construct is a static phenomenon used to enumerate all phonetic features which are typical of the given speechsound in its full, canonical form. A combination of some of these features then serves as a reference set for comparing one speechsound with another.

Such phonetic features are called inherent. For instance, in vowels we are talking especially about the open character of the vocal tract and vibration of the vocal folds from the articulatory viewpoint. From the acoustic viewpoint, the inherent features are the presence of fundamental frequency and formant structure. From the perceptual viewpoint, they include high sonority, vowel quality and quantity, and the percept of syllabicity. In plosives, the inherent phonetic features are the place of articulation, the occlusion (closure) and the release, their oral character, and the presence or lack of voicing.

Analyses of natural speech show that inherent features differ in their degree of **stability**. While the presence of some inherent features is nearly 100% even in highly reduced (implicit) pronunciation, other features tend to be weakened or completely dropped even in relatively cultured pronunciation. Table 1.3 shows several tendencies in the stability of phonetic features in Czech (Machač, 2004).

speechsounds	**more stable features**	**less stable features**
vowels	voicing, formant structure	quantity, quality, oral character
nasals	nasality	occlusion, place of articulation
voiced plosives	place of articulation, voicing	occlusion, presence of release
voiceless plosives	place of articulation, lack of voicing, occlusion	presence of release
voiceless sibilants	place and manner of articulation	

Table 1.3. Examples of the stability of inherent phonetic features in Czech.

1.3.2. Extrinsic phonetic features

Unlike the speechsound as a theoretical construct, its specific realization in speech takes place in time, it is a dynamic and highly variable phenomenon. Its physical qualities depend on diverse factors, among which the contrast with the neighbouring speechsound is the most important one for the purposes of this book.

Due to the smooth, continuous nature of articulation movements, as well as the way in which glottal and supraglottal gestures are synchronized, some phonetic features may extend beyond the boundaries of the speechsound to which they are inherent. This concerns, for instance, the tendency to the open character of the vocal tract, nasality, or the presence and absence of voicing. These features then co-determine the articulatory and acoustic character of neighbouring speechsounds, acting like non-inherent, extrinsic features. Naturally, this phenomenon can make segmentation considerably more difficult.

1.3.3. Segment boundaries and distribution of phonetic features

From the viewpoint of speech signal segmentation, a speechsound is, under ideal conditions (see Figure 1.3a), regarded as the time interval during which the inherent phonetic features (those characteristic for the given segment type) are present and, at the same time, these features are relevant for the differentiation from the neighbouring speechsounds. In this ideal situation, a segment boundary would be the time in which one speechsounds' inherent features would disappear and the other speechsounds' inherent features would emerge. It is obvious that such ideal situations rarely occur. A more realistic scenario of the distribution of phonetic features is depicted in Figure 1.3b which shows possible overlaps of selected features.

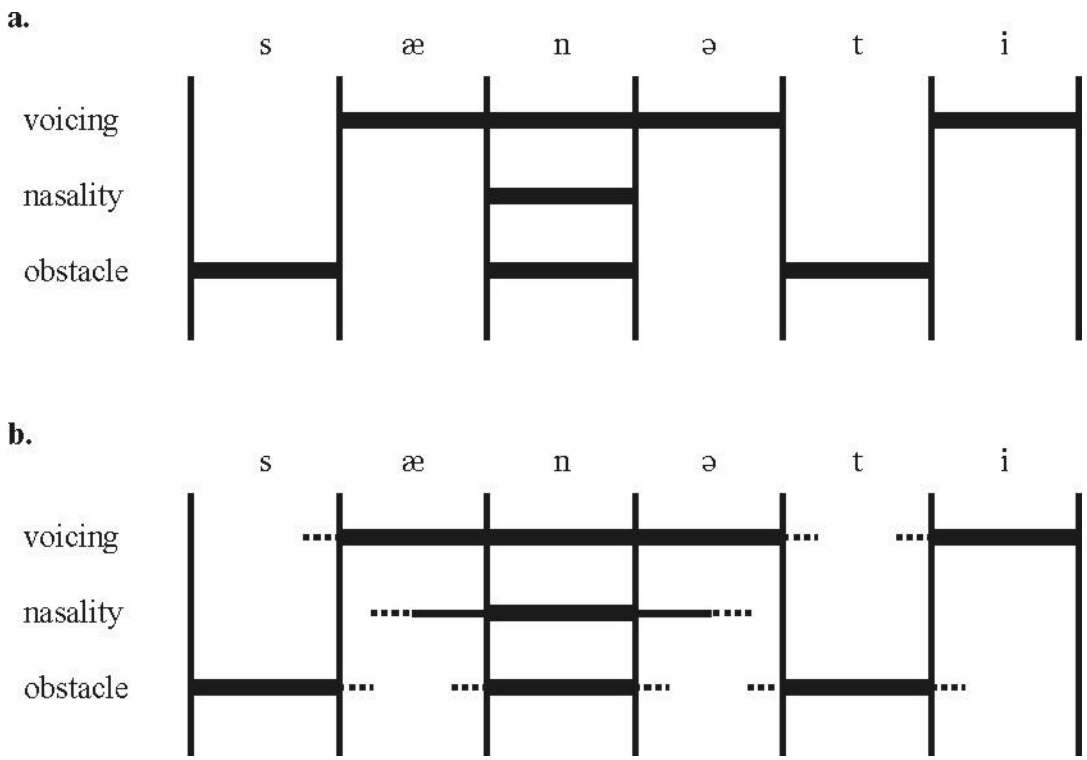


Figure 1.3. Schematic diagram of the distribution of phonetic features in the word *sanity*. **a.** Ideal distribution. **b.** Possible overlaps of features. Segment and overlap durations have been kept constant. (vertical lines = segment boundaries, thick horizontal line = inherent phonetic feature, thin horizontal line = overlap of a feature into the neighbouring segment, dots = onset and offset of a feature)

As we have mentioned above, one of the goals of our endeavour is to increase the accuracy of segmentation, i.e. the placement of boundaries in accordance with the rules, as allowed by the analyzed speech material. It is obvious, however, that segmentation accuracy depends mainly on the salience of acoustic contrast in the transition phase of the neighbouring segments. For cases when acoustic contrast is low, it is necessary to stipulate rules so that unclear items are dealt with as uniformly as possible.

1.4. Methodological and terminological remarks

In this section, we will present the first general segmentation rules, briefly introduce the material which we used to formulate the rules, and discuss several basic terminological and methodological aspects of segmentation.

First of all, let us stipulate three general rules:

1) As the presence of **formant structure** will be very important for segmenting many sequences, we try to place boundaries next to (or between) what we call **formant columns** (i.e., the dark vertical areas in the spectrogram, representing the peaks of acoustic energy in each glottal pulse). **Full formant structure** is characterized by a sequence of salient formant columns (especially in the area of vocalic F1-F3) - see Figure 1.4.

2) If there is a **transition phase** (an uncertain, "grey" portion of the signal in which low acoustic contrast does not allow unambiguous boundary placement according to the rules), the boundary will be placed in the temporal midpoint of this area (or, more precisely, at the nearest zero crossing from the midpoint). By using the midpoint rule whenever necessary, we try to prevent as much as possible systematic or random differences which may affect phonetic analyses - one item with a transition phase and its marking may be found, for example, in Figure 2.10 (page XX).

3) Boundaries will be placed at a **zero crossing** (a point in which the waveform crosses the amplitude axis) - see Figure 1.4. Naturally, this rule may be abandoned in microstructure investigations where every single millisecond counts.

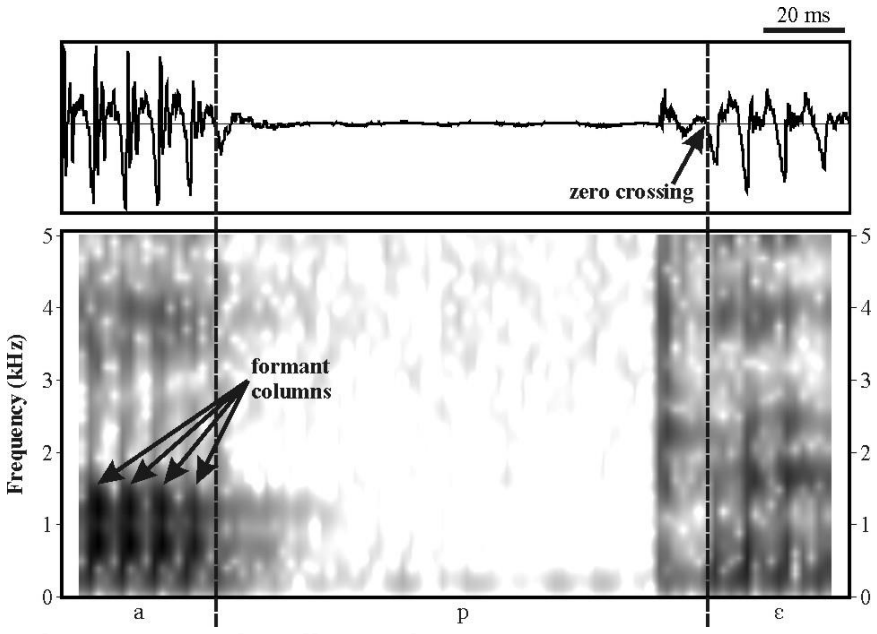


Figure 1.4. A prototypical example of illustrations used in this handbook, with the arrows pointing to a zero crossing and to formant columns.

Throughout the book, we will be discussing various kinds of acoustic contrast between neighbouring speechsounds. We would like to draw the reader's attention to an important contrast between **intensity** and **amplitude**. By intensity, we mean spectral intensity depicted in the spectrogram by different shades of grey. Amplitude, on the other hand, will refer to the waveform, specifically to the maximum amplitude or to changes in amplitude from one period to another. This contrast is important because, as can be seen in Figure 4.2, relative spectral intensity and peak amplitude need not correlate.

All segmentation examples in this handbook have been created in Praat (Boersma & Weenink, 2009) and graphically processed in CorelDRAW X3. The illustrations used in this handbook typically present both the waveform and the spectrogram. The spectrogram is used by itself only when the waveform does not aid differentiating between the given segment types. Following the default setting in Praat, the spectrograms display the frequency range from 0 to 5 kHz; in the chapter on fricatives, however, the 0-8 kHz range is displayed. When segmenting connected speech, we therefore recommend the labeller to display the 0-8 kHz range; this may facilitate the segmentation of not only fricatives but also other speechsounds. Time is not labelled in the illustrations, but a 20-millisecond scale is indicated at the top of each example. Figure 1.4 illustrates the features and graphical conventions described so far.

Several types of speech material were used to formulate these segmentation guidelines:

a) Czech university students - recordings of read speech (75 students were asked to read a short story after a few minutes of preparation),

b) Czech university students - semi-spontaneous recordings (the same 75 students were asked to tell a story based on picture prompts),

c) Czech radio - recordings of news bulletins read by six newsreaders,

d) BBC World Service - for examples from English, we used recordings of BBC news bulletins read by 12 newsreaders.

There is one final point which remains to be mentioned before we introduce the structure of the book. So far, we have talked about boundaries whose location is relatively easy to determine, as well as about boundaries placed near the midpoint of a definable transition area. That might give the impression that speechsound boundaries can always be located visually, which is, obviously, not the case. Listening is often necessary, if only to confirm the visual cues. Listening will actually quite often serve as the primary cue for determining the boundary between two speechsounds. In such cases, listening will mean "careful listening" in which we try shifting the boundary and check the auditory impression (*cf.* Kvale & Foldvik, 1991).

In the following chapters, we present guidelines for boundary placement in selected speechsound combinations. The chapters start with a brief summary of the articulation and acoustic properties of the target segment or segment type. Next, we introduce the relevant phonetic features for the given segment type and, based on those, formulate segmentation rules for both canonical and less explicit

pronunciation. Each chapter ends with a short recapitulation of the main segmentation principles. Chapters 2 to 7 present guidelines for segmenting intervocalic consonants, Chapters 8 to 10 deal with boundaries between members of consonant clusters, and Chapters 11 and 12 examine issues related to word- and utterance-initial, as well as utterance-final contexts.

2. Intervocalic plosives

2.1. Articulatory and acoustic lead-in

Plosives are obstruent consonants whose articulation consists in two main stages: the **occlusion** (or **closure**) and the **plosion** (or **release**). During the occlusion, the vocal tract is closed completely so that air is accumulating before the closure (in the direction of airflow from the lungs) and intraoral pressure is rising. During the **plosion** phase, the closure is released, the fast flow of air between the articulators causes the balancing of intraoral and atmospheric pressure, and short turbulent noise is generated behind the closure (in the direction of airflow from the lungs).

Acoustic characteristics in the two stages of plosives differ somewhat depending on their voicing status. In **voiceless plosives**, there is no fundamental frequency (F0) nor formant structure present in the occlusion phase. Since glottal and supraglottal activity are not completely synchronized, we can often see voicing continuation (Stevens, 1998: 333) at the beginning of the closure for some 20-30 ms. The burst of the plosion tends to be short and salient in voiceless plosives. In (fully) **voiced plosives**, F0 is present throughout the closure. The plosion tends to be quite weak, with the aperiodic noise component often completely missing.

All the acoustic events in plosives (occlusion, release, voice onset time, transitions to neighbouring vowels) tend to differ in duration depending on the place of articulation as a result of different mass of the articulating organs which are required to complete the given gesture, as well as the organs' mobility (Machač, 2006). Apart from relative duration, plosives (especially voiceless plosives) also differ in the spectral composition of the release noise, with alveolar burst being strongest in frequency (around 4-5 kHz), velar burst lying around 1.5-2 kHz (though velars tend to be strongly affected by the quality of the following vowel), and bilabials having a relatively flat spectrum.

2.2. Inherent phonetic features and basic segmentation rules

The inherent phonetic features of plosive consonants are: a) complete closure at the place of articulation and the consequent absence of formant structure, b) presence of the release in the form of a noise burst, c) oral character, and d) presence of fundamental frequency (F0) in phonologically voiced plosives, and absence of fundamental frequency in phonologically voiceless plosives.

When formulating segmentation rules, we try to find such inherent phonetic features of both types of speechsounds (i.e., plosives and vowels) which are relevant for their differentiation in the acoustic signal. Some features, though important as characteristics of speechsound types, cannot serve as reliable and practical cues for boundary placement. First, let us have a look at features

which are not suitable for segmentation of intervocalic plosives: fundamental frequency and the noise burst.

Obviously, using the presence or absence of F0 is problematic because it is a feature common to vowels and voiced plosives. If we were to use F0 for finding the boundaries of voiceless plosives, we would face two problems: first, voicing continuation from vowels into the "voiceless" occlusion and second, comparability between segmenting voiceless and voiced plosives.

As for exploiting the noise burst for locating the right boundary of plosives, the problem consists in the fact that the plosion tends to be considerably weaker and shorter in voiced plosives than in voiceless plosives, and its end can be difficult to detect. Comparability of segmenting voiced and voiceless plosives would again be difficult and the location of the boundary would not completely correspond to actual changes in the vocal tract.

The features which should be exploited for the segmentation of intervocalic plosives are, then, the open character of the vocal tract and thus **presence of full formant structure** in vowels, and the closed character of the vocal tract and thus **absence of formant structure** in plosives.

Figure 2.1 shows the course of selected phonetic features in intervocalic plosives: fundamental frequency, F1 and F2, and noise. Black colour indicates the highest degree of presence of the given feature, progressively lighter and darker shades of grey express, respectively, weakening and strengthening of the feature, and white colour indicates absence of the feature.

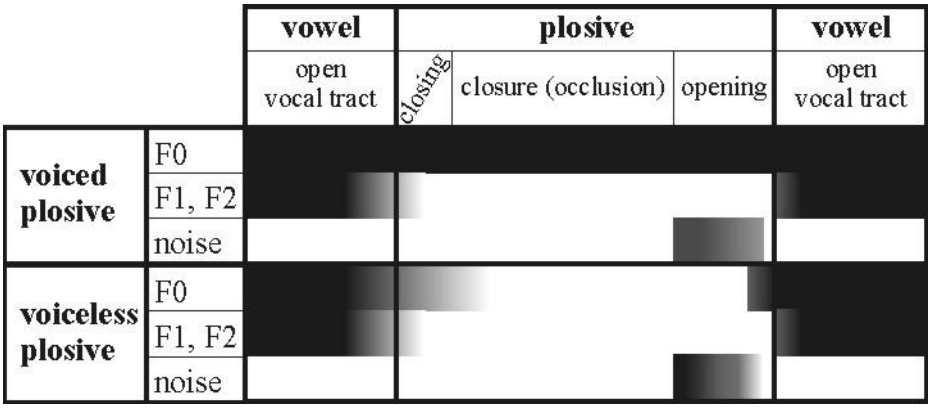


Figure 2.1. A schematic diagram of the course of F0, F1 and F2, and noise in a vowel-plosive-vowel sequence.

To summarize, the different character of voiced and voiceless plosives does not allow comparable segmentation based only on their own articulatory and acoustic properties. That is why the contrast between the presence and absence of **full formant structure** will be the inherent feature on which we will base our segmentation decisions.

In Figure 2.2, the sequence [ucɪ] is divided into six acoustic events (indicated by the numbers between the waveform and spectrogram), and speechsound boundaries based on the "full formant structure" criterion are given. Again, we can see how misleading it would be to exploit fundamental

frequency for segmentation of plosives. Figure 2.3 shows another example of an intervocalic plosive, this time with a (visually) less salient noise burst of the bilabial [p].

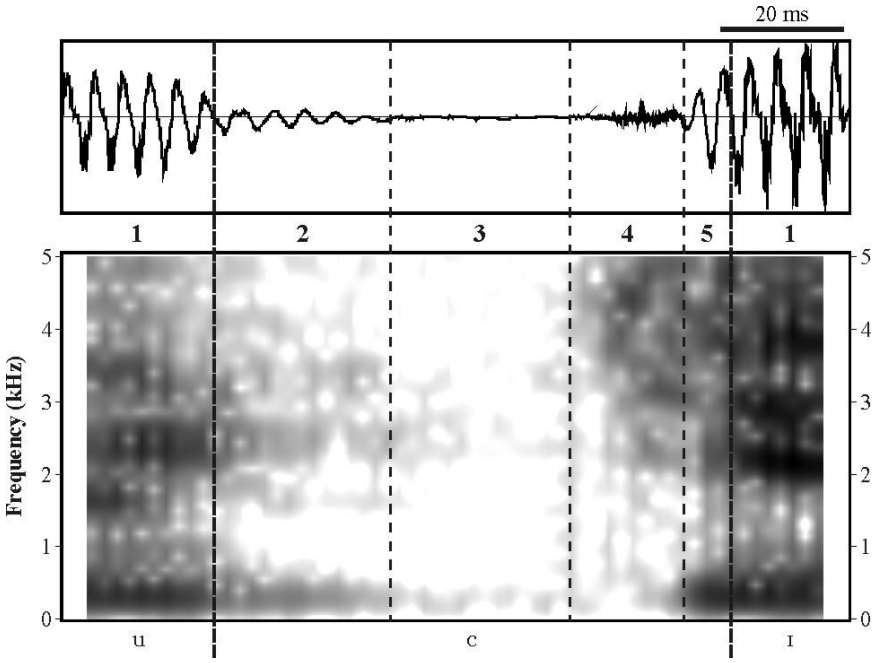


Figure 2.2. Sequence [ucɪ] illustrating the application of the full formant structure criterion. The acoustic events indicated are full formant structure of the vowel (1), voicing continuation in the closure (2), closure without voicing (3), noise (4), and F0 onset (5). The plosive is regarded as a succession of stages 2-5.

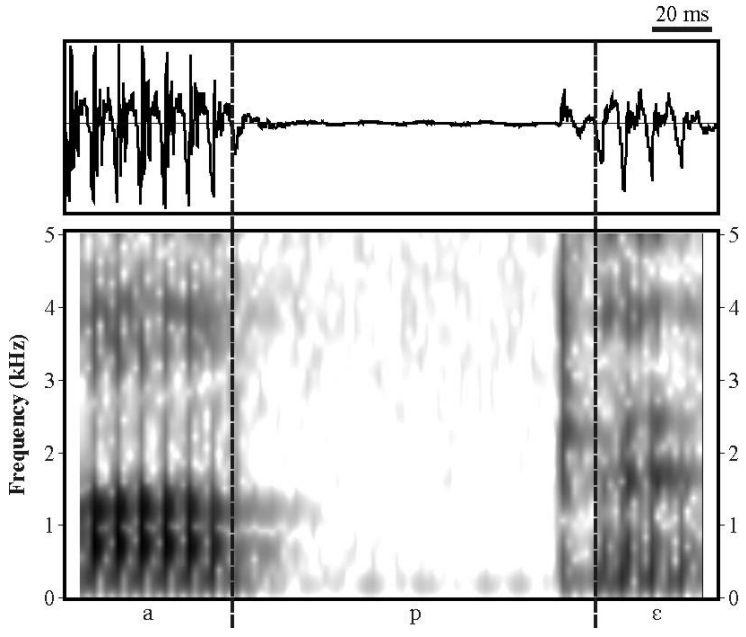


Figure 2.3. Sequence [apɛ] with a less salient plosion.

It has been mentioned above that the release in voiced plosives tends to be weak and sometimes with only barely visible noise in the waveform, even though the speaker produces a complete closure. Figure 2.4 shows such an example in the sequence [abɪ]; the onset of full formant structure is well visible, though, and segmentation is not problematic here.

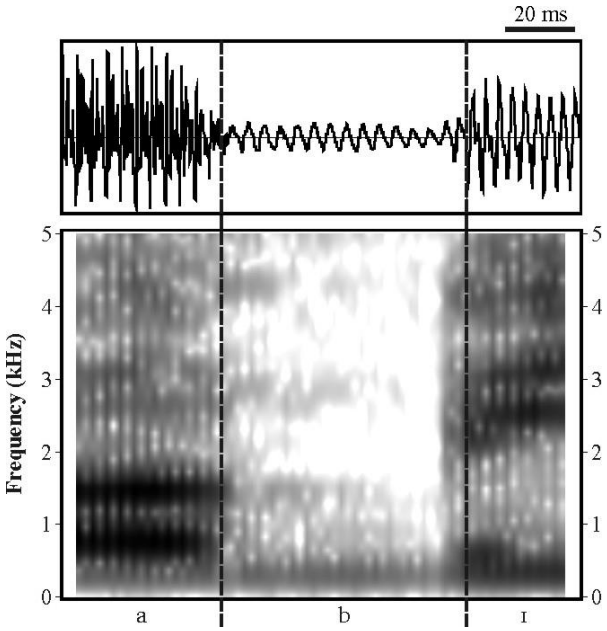


Figure 2.4. Sequence [abɪ] with a weak plosion, almost without noise components.

2.3. Additional segmentation guidelines

This section will introduce examples in which segmentation is less straightforward, the main reason being that the plosives were not pronounced in the canonical manner. One of the inherent phonetic features of plosives which may not be present is complete closure. With some significant narrowing in the vocal tract preserved, we can talk about spirantization of the plosive. Segmentation is then proceeds similarly as in fricatives (see Chapter 3 for more detail), but we can say here that the onset of full formant structure is, again, the primary cue, as indicated in Figure 2.5 and especially Figure 2.6. Although both these figures show spirantization of the palatal [ɟ], this phenomenon is not limited to this context.

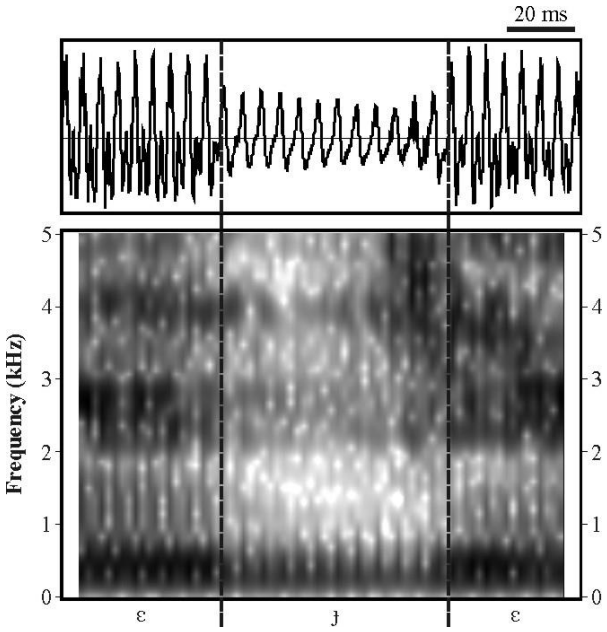


Figure 2.5. Sequence [ɛɟɛ] with weak spirantization.

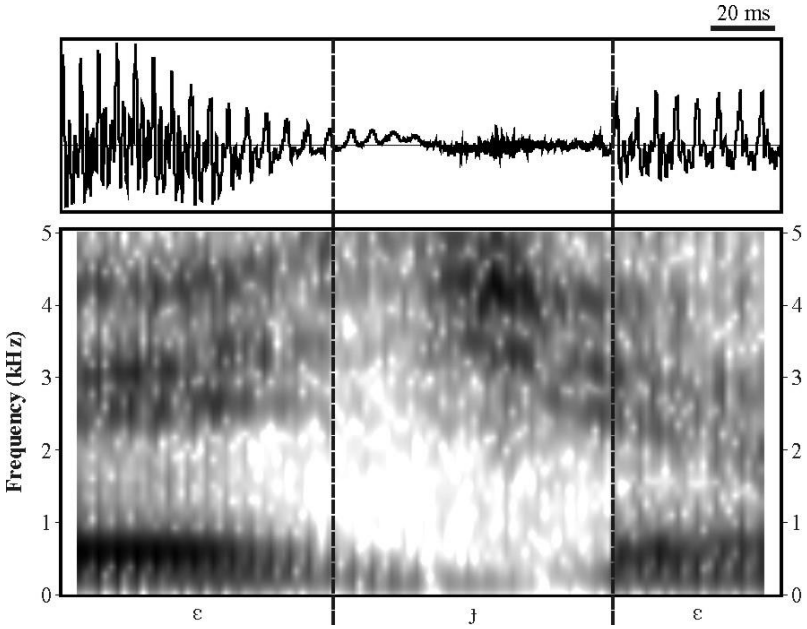


Figure 2.6. Sequence [ɛɟɛ] with stronger spirantization (and also partial devoicing in the closure).

More problematic instances are those in which a voiced intervocalic plosive obtains, due to the absence of complete closure, a sort of "semivowel" character. Figure 2.7 shows another sequence [ɛɟɛ] where the plosion is not visible at all. Due to the semivowel character of [ɟ], the onset of formant structure is somewhat more gradual, and boundary location is thus a bit more difficult. The following cues in the signal usually have to be taken into account: the salience of formant structure, the complexity of the waveform shape, relative amplitude, as well as perceptual impression.

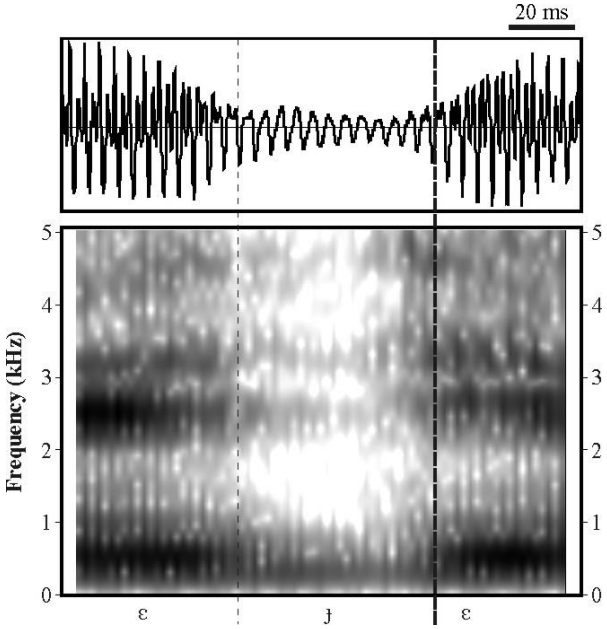


Figure 2.7. Sequence [ɛɟɛ] with a semivowel-like /ɟ/.

Especially velar plosives often display multiple plosions. We are mentioning this interesting phenomenon, although multiple plosions do not constitute a problem from the segmentation

viewpoint, because we are interested in the onset of full formant structure. Figure 2.8 shows an example of [k] with three bursts.

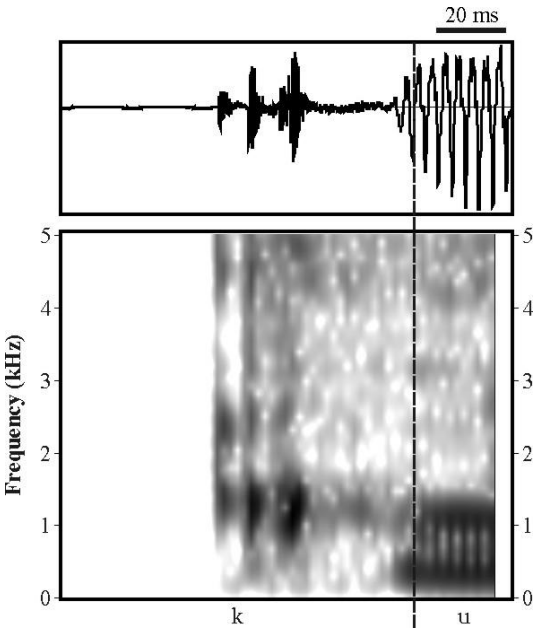


Figure 2.8. Sequence [ku] with a multiple plosion.

We have mentioned in the Introduction that it is impossible in certain situations to determine an exact time of onset or offset of a specific acoustic phenomenon, and that there may be shorter or longer phases of gradual transition, i.e., **transition phases**. In those instances, inherent phonetic features which are considered relevant for the differentiation between neighbouring speechsounds are not salient enough in the acoustic signal to allow unambiguous boundary location, because acoustic contrast in the transition phase is low. The transition phase can assume several forms, for example:

a) the decay of formant structure is gradual, it can last as long as several tens of milliseconds,
b) formant columns are not salient,
c) a noise component interferes with the formant structure.

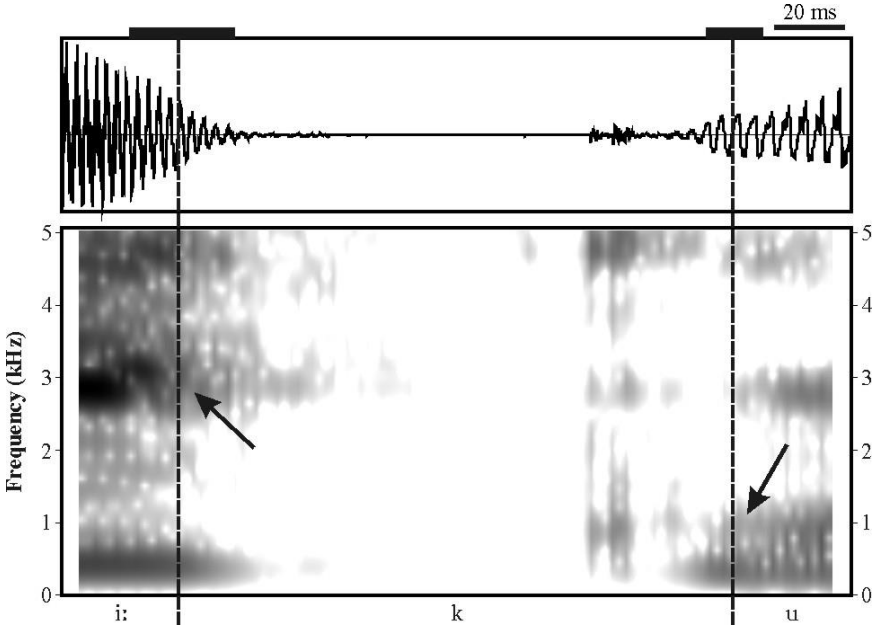
In plosives, transition phases are more frequent in the vowel-plosive sequence, since the approximation of the articulating organs (the formation of the closure) is slower than the opening (which is driven by the accumulated air pressure).

In Figure 2.9, we can see a transition phase on both sides of the plosive. In both instances, the boundary is placed in the midpoint of this transition. In the sequence [iːk], the transition phase consists in the very slow decay of formant structure (indicated by the black bar). In the sequence [ku], the onset of formant structure is also quite slow. Figure 2.10 shows another example of slowly decaying formant structure, especially in the F2 region slightly below 1 kHz.

Interesting transitions seem to occur in sequences of [iː] with a dorsal (i.e., velar or palatal) plosive. The closing gesture for the plosive in these sequences often starts during the articulation of the vowel, which leads to the superposition of friction, similar to a voiced palatal fricative [ʝ], over the

vowel quality. The decision as to how to deal with such instances appears to be rather arbitrary. One could put the boundary before or after the noise; we have chosen to consider the noise as the transition phase, and therefore to place the boundary in the middle of the noise (see Figure 2.11).

Figure 2.9. Sequence [iːku] with relatively long transition phases on both sides of [k]. The horizontal bars indicate the extent of formant structure decay and onset, the arrows indicate their location along the frequency scale.



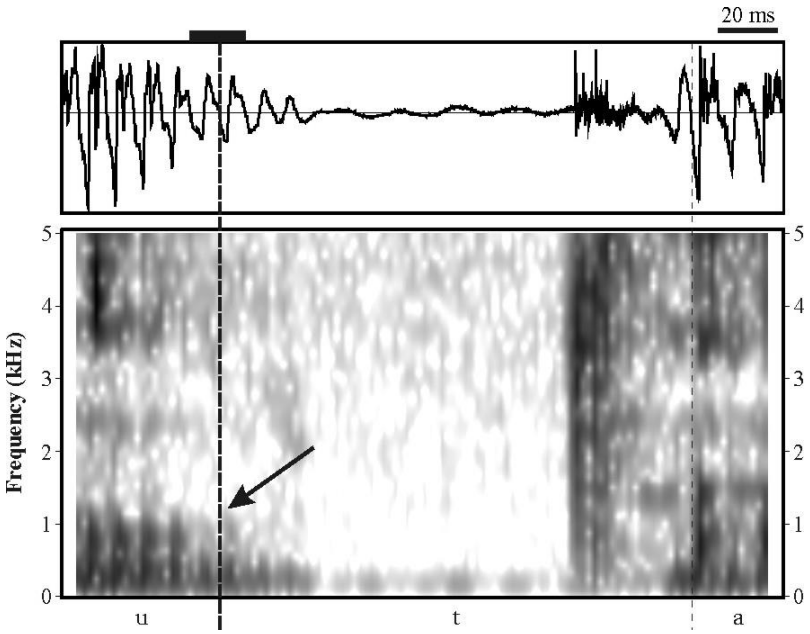
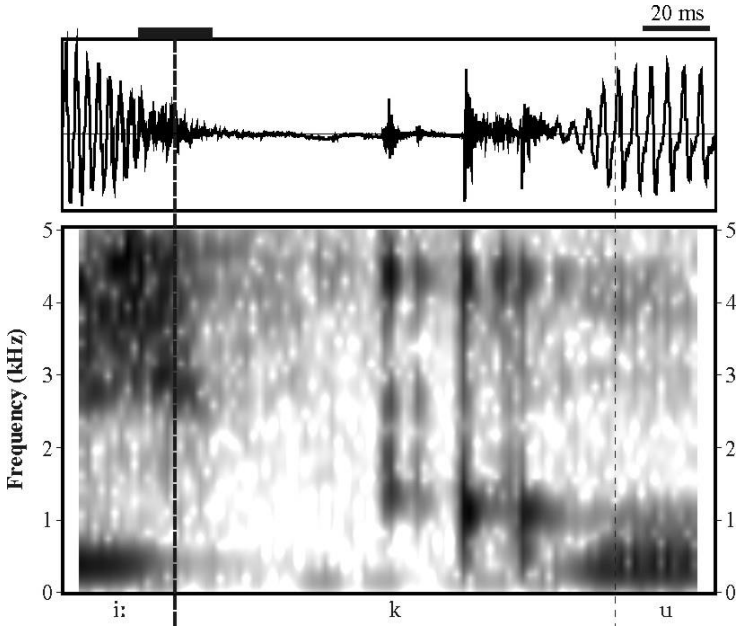


Figure 2.10. Sequence [uta] with slowly decaying formant structure. The horizontal bar indicates the extent of formant structure decay, the arrow indicates its location along the frequency scale.

The last phenomenon which remains to be mentioned with respect to plosives is the frequent implicit, or "short" pronunciation of alveolar /d/. Instead of achieving a complete closure, the tongue tip performs merely a ballistic movement, resulting in what is called a **flap**. The alveolar flap [ɾ] is very frequent in American English as a variant of both alveolar plosives, /t/ and /d/;

words like *latter* and *ladder* then become essentially homonymous, ['lærɚ]. In Czech, [ɾ] is also frequent as a variant of the trill /r/ (see Chapter 5), and it is the standard pronunciation of the



"short" *r*-sound in Spanish.

Figure 2.11. Sequence [iːku] with palatal noise. The horizontal bar indicates the transition phase.

Figure 2.12 shows two versions of flap realization of /d/ in the Czech word *budu*. In the item on the left, the formant structure is weakened in higher frequencies as a result of the very short approximation of the tongue tip to the alveolar region. In the item on the right, the right boundary of this speechsound is more difficult to determine only visually, and listening must be used to aid segmentation.

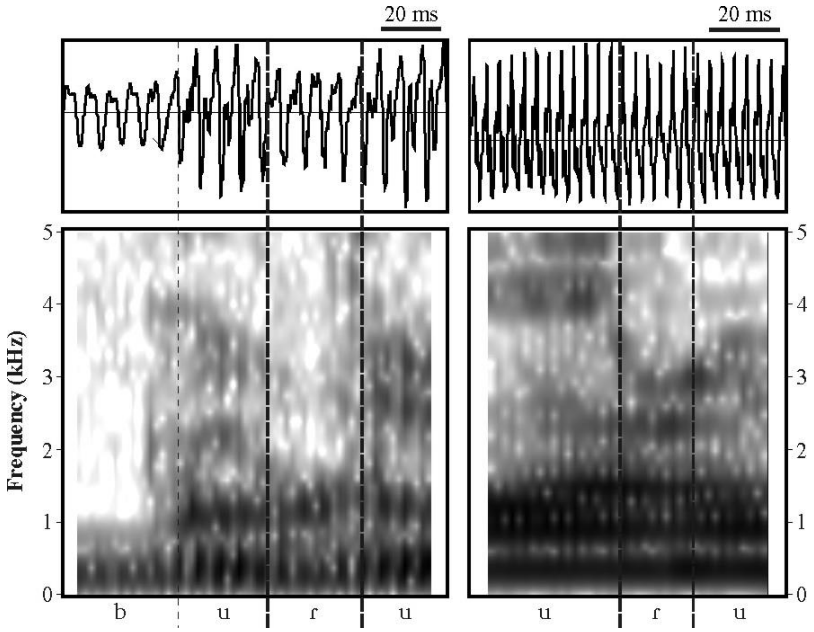


Figure 2.12. Two sequences of [ɹu], the one on the left being more explicit than the one on the right.

2.4. Summary

When locating boundaries in sequences with canonically pronounced plosives, we apply the full formant structure criterion. In other words, the plosive starts at the offset of full formant structure (at the end of the last salient formant column) and ends with the onset of full formant structure (at the beginning of the first salient formant column). Voicing continuation at the beginning of the closure, as well as F0 onset at the end of the plosion are regarded here as part of the plosives.

If the decay or onset of formant structure is gradual or indistinct, we regard that portion of the acoustic signal as a transition phase and place the boundary of the plosive into the middle of that area (that is, at the zero crossing nearest to the midpoint of the transition phase).

In items without complete closure, we exploit mainly the salience of formant structure in the neighbouring vowels, and then the shape of the waveform, amplitude drop and, when necessary, listening. More detail on segmenting plosives with a spirantized and semivowel character (see above) can be found in chapters on segmenting fricatives and approximants, respectively.

Finally, let us mention aspirated plosives. Aspiration is frequent in fortis plosives which lie in stressed syllables in English or German. Since aspiration is acoustically similar to voiceless fricatives, the right boundary of aspirated plosives will be segmented as that of voiceless fricatives (see Chapter 3).

3. Intervocalic fricatives

3.1. Articulatory and acoustic lead-in

The articulation of fricatives consists in forming a **stricture**, in the form of a critical **narrowing** of the airstream passage at some place of the vocal tract. Airflow accelerates in the area of the narrowing, which leads to the generation of turbulence behind the narrowing. It is this turbulence which we perceive as fricative noise. The spectral properties of the noise depend mainly on the place of articulation and, therefore, on volume ratios of resonance cavities. Fricatives are the only manner of speechsounds which, in the languages of the world, exploits all the common places of articulation within the vocal tract, from the lips to the glottis.

The narrowing is typically formed along the central line of the vocal tract. In most lingual fricatives, for instance, the sides of the tongue are pressed to the roof of the mouth, and a part of the tongue between its sides approaches the roof of the mouth at some place, so that a sufficient (critical) narrowing is achieved for the acceleration of airflow. The opposite situation occurs in lateral fricatives [ɬ ɮ], where the tongue touches the central portion of the alveolar ridge, while the sides of

the tongue are lowered so that a stricture is formed between them and the sides of the hard palate (*cf.* section 7.1 for the lateral alveolar approximant [l]).

In **voiceless fricatives**, the glottis is open and F0 is therefore missing. Voiceless fricatives are displayed as a continuous area with higher intensity in a relatively broad frequency band (the so-called noise formant), whose position is determined by the place of articulation. In other words, we are talking about noise with no periodic component. The waveform shows only aperiodic noise. At the beginning and end of voiceless fricatives, we may see voicing continuation and voicing onset, respectively (*cf.* parallel phenomena in plosives, section 2.1).

In (fully) **voiced fricatives**, both noise and periodic components should be present throughout the articulation of the speechsound. The two components are generated at different places in the vocal tract: the noise in and behind the narrowing, the periodic component at the glottis. A canonically formed voiced fricative is displayed in the spectrogram as a relatively broad frequency band, with vertical striations and the voice bar indicating periodicity. Aerodynamic relations during the production of voiced fricatives are more complicated than in voiceless fricatives, because it is quite difficult to keep different pressure at two pressure spots (the vibrating vocal folds and the stricture). That is why one can often encounter either devoicing (partial or complete) or loss of friction in phonologically voiced fricatives in the intervocalic position.

We can draw a parallel between the spectral composition of fricative noise and plosive release bursts. As in plosives, the noise formant is highest in alveolar fricatives (around 4-5 kHz), somewhat lower in palatal (3-4 kHz) and post-alveolar fricatives (2.5-3.5 kHz), and still lower in velar fricatives (1-1.5 kHz). Labial fricatives, which do not have any specific filtering, have relatively flat spectra. Since the strongest components only start appearing at about 4 kHz in alveolar fricatives, it might help in some less straightforward situations to display a greater range of frequencies. Although it is not essential in all items, the spectrograms in this chapter will all display the 0-8 kHz frequency range.

3.2. Inherent phonetic features and basic segmentation rules

The inherent phonetic features of fricative consonants are: a) critical narrowing, or stricture at the place of articulation, which is reflected in the presence of noise components, and b) presence of fundamental frequency in phonologically voiced fricatives, and absence of fundamental frequency in phonologically voiceless fricatives.

From the viewpoint of segmenting intervocalic fricatives, we consider the **onset and offset of full formant structure** in the vowel to be the decisive factor, rather than the **presence of fricative noise** - at least in canonical sequences. The reason for this decision is, as it will be as much as possible in subsequent chapters, comparability of segmentation of various speechsound types. In other words, the presence of formant columns (see section 1.4) in the spectrogram will be regarded

as more important than the “spiky” or “hairy” character of the waveform, or the presence of a broadband maximum in the spectrogram.

The application of this rule is illustrated in Figure 3.1 for a voiceless fricative. As in plosives, we have to distinguish between full-fledged formant structure and F0 continuation or prevoicing, as indicated in the figure by the arrows. Due to incomplete synchronization of glottal and supraglottal activity, F0 continuation and prevoicing may last one or two periods.

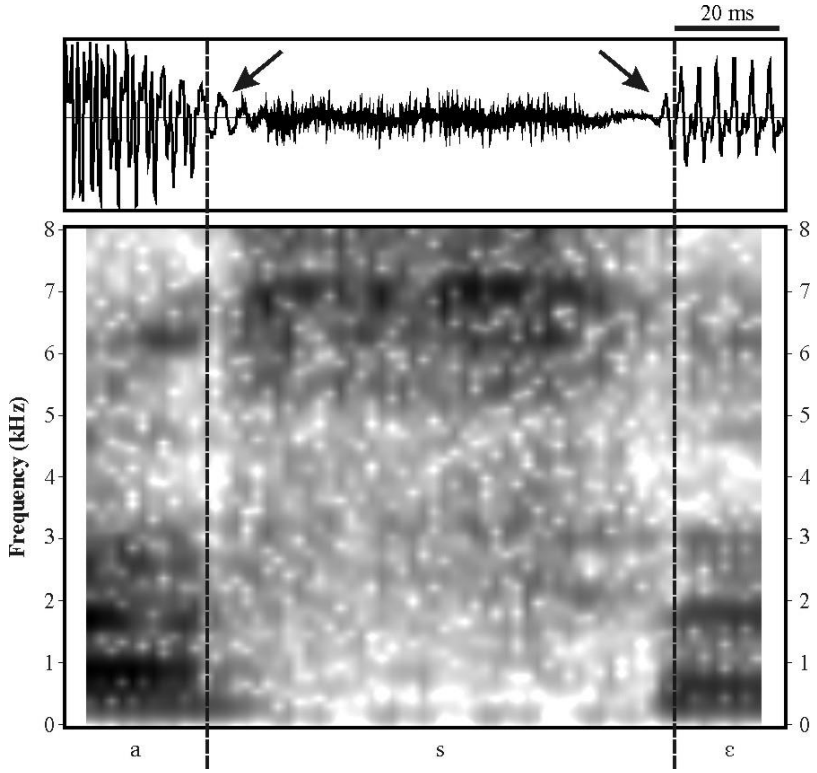


Figure 3.1. Sequence [asɛ] showing the segmentation of a canonical voiceless intervocalic fricative based on the vowels’ full formant structure.

As a secondary feature, we can exploit **relative intensity differences** (in programs which enable the visualization of the intensity contour). Typically, intensity drops in vowel-fricative sequences and increases in fricative-vowel sequences. The boundary, located using the inherent features, often coincides with approximately the midpoint of the intensity drop or increase (see Figure 3.2 for an example with a voiced fricative).

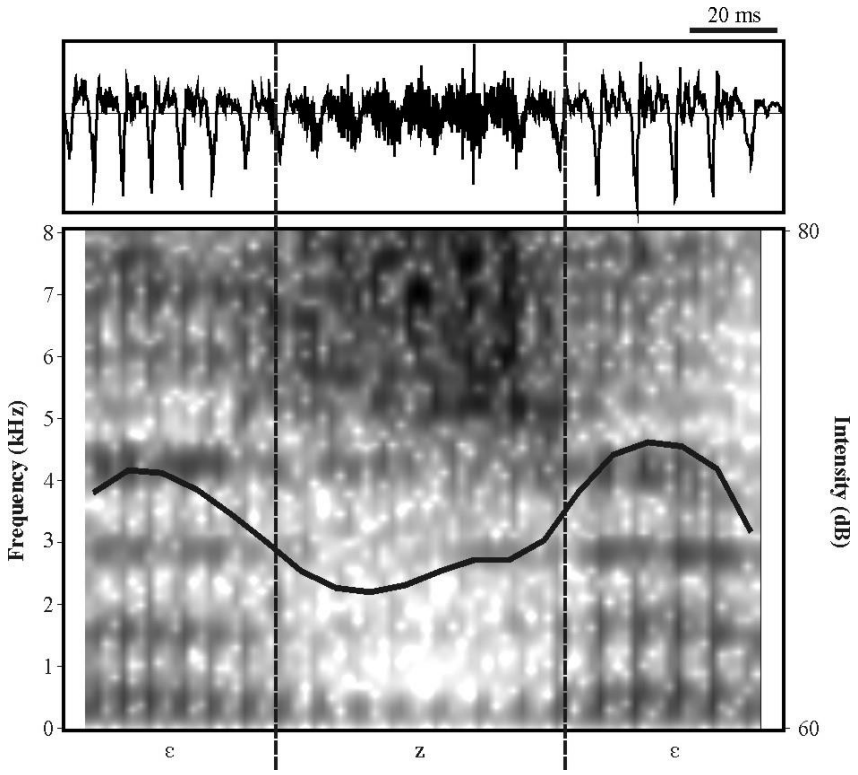


Figure 3.2. Sequence [ɛzɛ] showing the segmentation of a canonical voiced intervocalic fricative. The thick curve corresponds to the intensity contour and is related to the scale on the right.

3.3. Additional segmentation guidelines

Typically, the location of the boundary is more difficult to determine in vowel-fricative sequences. This is caused by the tendency for the decay of formant structure to be more gradual than the onset of formant structure in fricative-vowel sequences. That is why it is likely for the **midpoint of the transition area** rule to be exploited quite often when we segment intervocalic fricatives, especially in items where pronunciation is less careful. The transition is identified in these instances as the portion of the signal between the beginning of the decay of full formant structure and the beginning of full-fledged fricative noise and vice versa, as illustrated in Figure 3.3.

So far, the illustrations in this chapter featured exclusively alveolar and post-alveolar fricatives. Apart from their high frequency, the reason is also their relatively salient spectra (prominent maxima in the spectrum). Spectral salience also applies to the voiceless velar fricative [x]. Figure 3.4 shows a sequence with intervocalic [f]; as we have mentioned in the first section of this chapter, labiodental fricatives tend to have quite flat spectra. However, this usually does not constitute a major problem for segmentation.

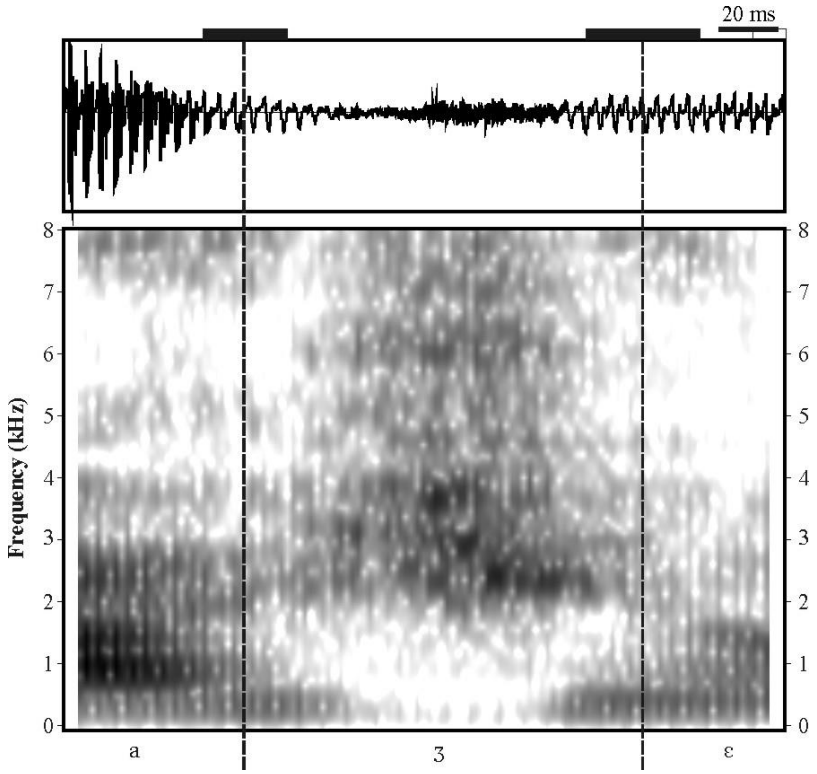


Figure 3.3. Sequence [aʒɛ] illustrating the slow decay and onset of formant structure. Boundaries are placed near the midpoint of the transition area (indicated by the horizontal bars at the top of the figure).

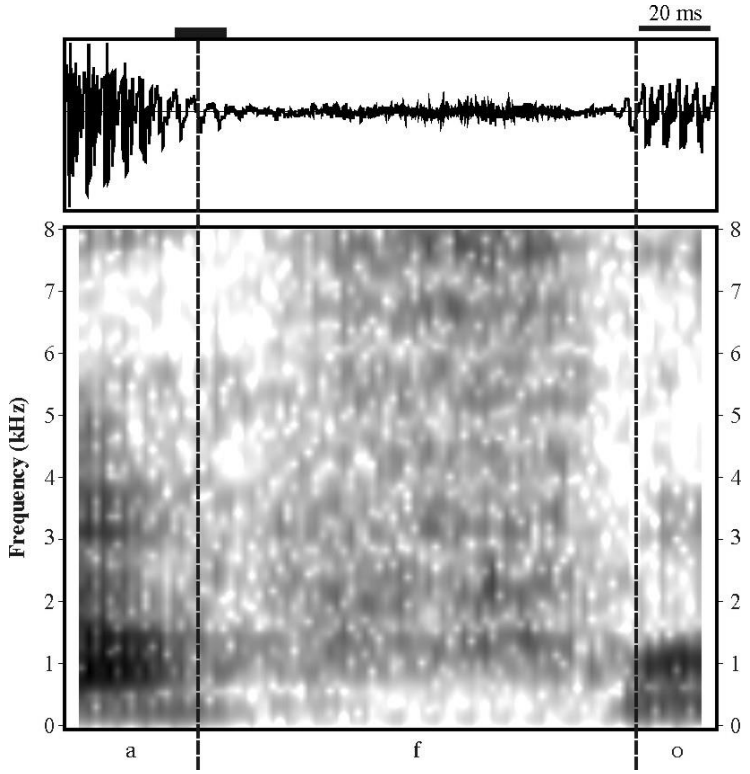


Figure 3.4. Sequence [afo] with the flat spectrum of the labiodental fricative.

There are two more sounds classified as fricatives and frequent in languages of the world which, however, have somewhat different acoustic properties, and the above-mentioned segmentation

guidelines thus need not always apply. The first one is the voiced labiodental /v/, the other is the laryngeal /h/. They will be analyzed in more detail in the next section.

3.4. The "less fricative" fricatives, /v/ and /h/

The first of these speechsounds, /v/, is described as a fricative but often loses friction in the intervocalic position and becomes more of a labiodental approximant, [ʊ]. This has been documented before for Czech (Skarnitzl & Volín, 2005) but seems to occur in English, too. Figure 3.5 shows an item of intervocalic /v/ with friction from the English word *avoid*; the fricative character of /v/ may partly be caused by the fact that it appears at the beginning of a stressed syllable.

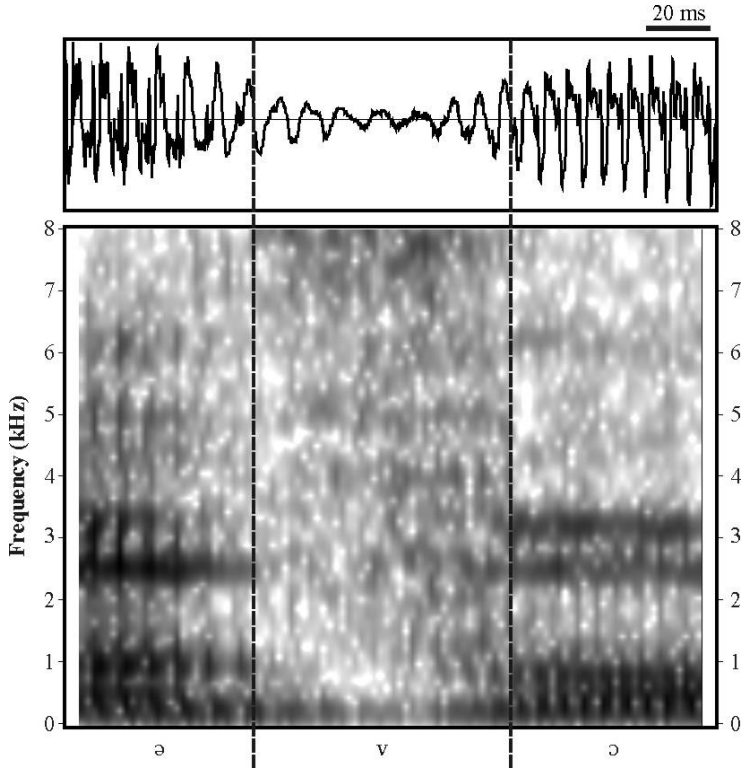


Figure 3.5. Sequence [ə'vɔ] showing the segmentation of /v/ with a substantial noise component and thus high acoustic contrast with the neighbouring vowels.

Figure 3.6 shows the sequence /ɪvɪ/ from the English word *giving*, and we can see that there is almost no fricative noise identifiable in the signal, making the speechsound more similar to an approximant. In this particular item, relative intensity of formants is sufficient for comparatively straightforward segmentation.

In Czech, intervocalic /v/ is hardly ever realized so explicitly that the location of boundaries is clearly visible in the signal. The presence of a noise component can be regarded as rather exceptional. What is therefore important for segmentation is **changes in formant structure**, possibly **changes in overall intensity and waveform shape**. Quite frequently, it is necessary to use listening. Figures 3.7 and 3.8 show two examples with no friction in the consonant. It is possible in both of them to exploit, at least to an extent, relative intensity and amplitude differences. The cues

available in the waveform and in the spectrogram are quite similar to those described in Chapters 6 and 7, in which we deal with glides and with the lateral approximant /l/, respectively.

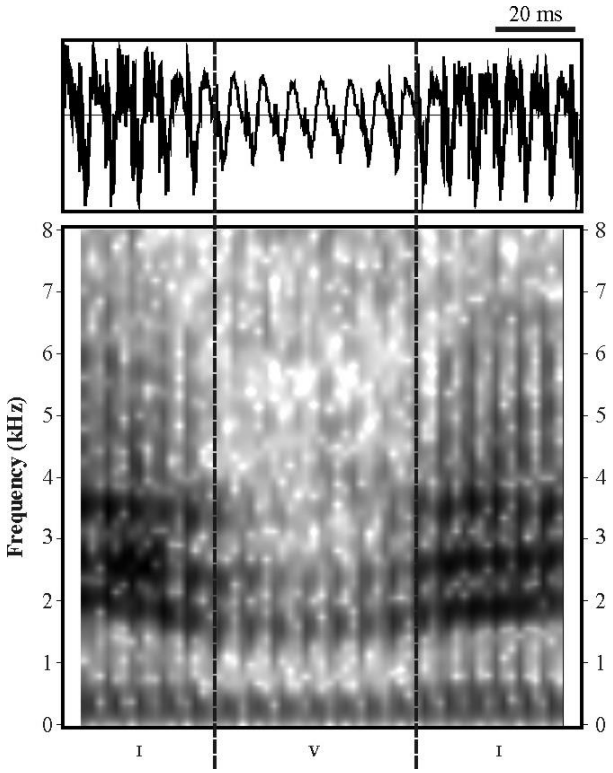


Figure 3.6. Sequence [ɪvɪ] showing an approximant character of English /v/; segmentation is relatively easy thanks to salient differences in formant intensity.

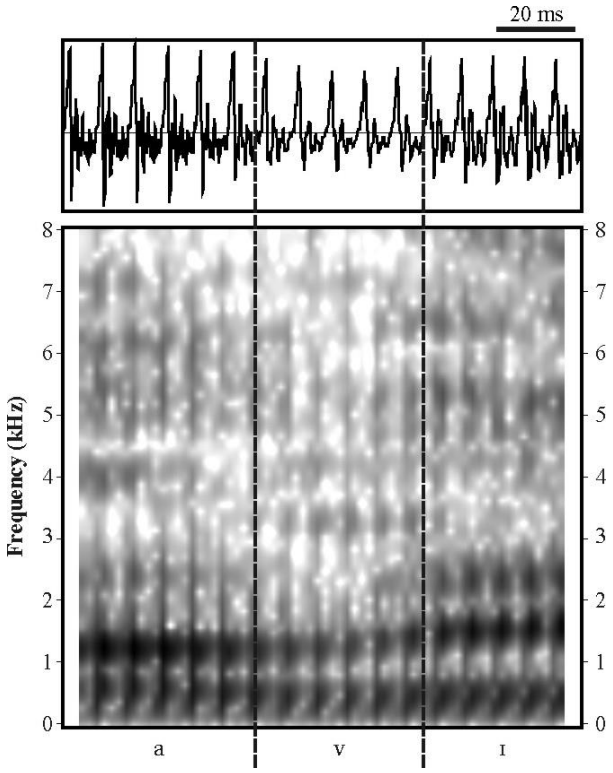


Figure 3.7. Sequence [avɪ] with an approximant-like /v/. Relative formant intensity, energy in high frequencies, as well as slightly lower amplitude in the waveform may be exploited for the location of the boundary. Listening facilitates segmentation, too.

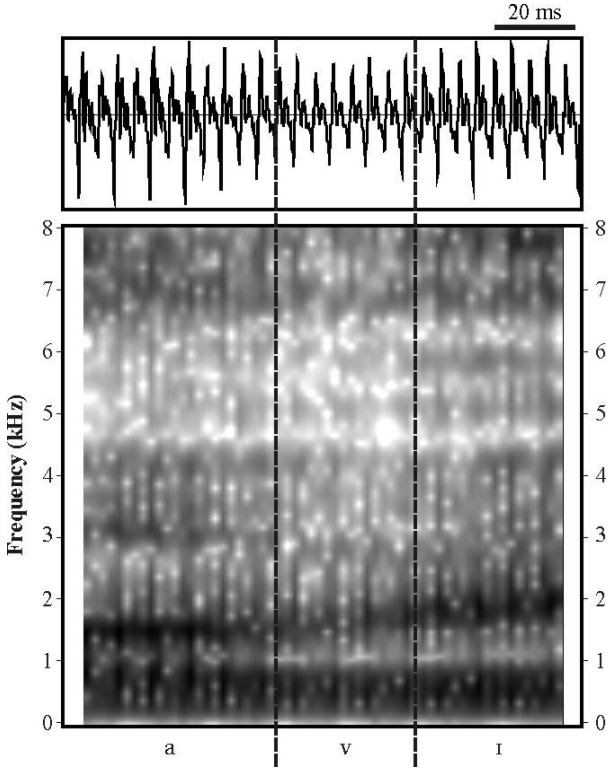


Figure 3.8. Sequence [avɪ] with an approximant-like /v/. Listening typically must be used in items like this one, at least to confirm the visual cues.

The second of the "less fricative" fricatives can have two forms: voiceless [h] and voiced [ɦ]. The voiceless variant does not pose new challenges for segmentation; acoustic contrast tends to be high, although it may be necessary to apply the rule placing the boundary near the midpoint of the transition area (see Figure 3.9, which shows the sequence [ɪˈhə͡ʊ] from the word *behold*).

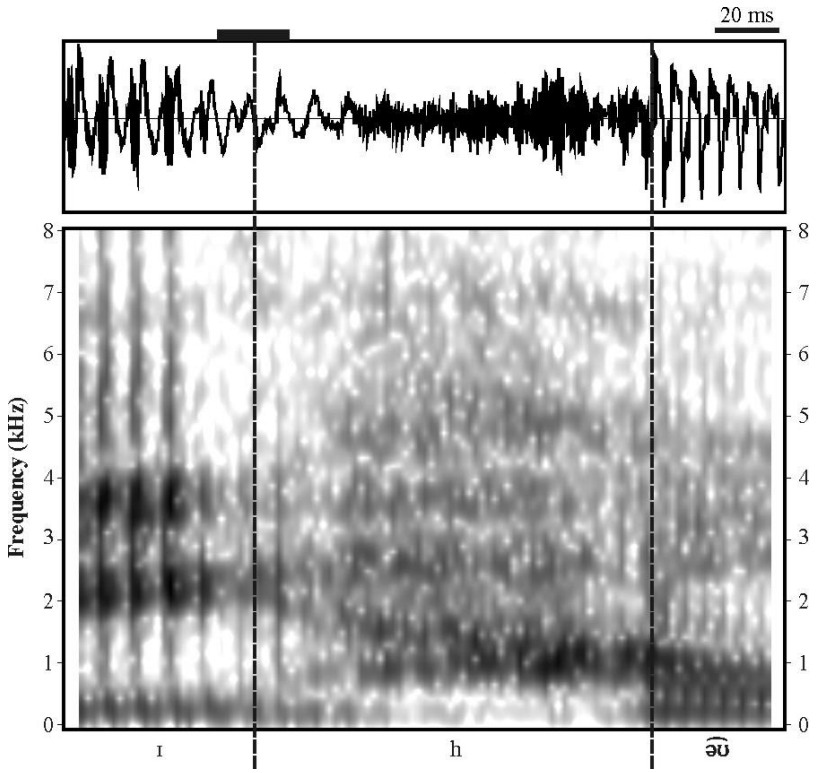


Figure 3.9. Sequence [ɪˈhəʊ] with the transition area between [ɪh] indicated by the black bar.

Czech has the rather rare voiced /ɦ/. Laryngeal fricatives in general do not have fricative formants of their own. Formants are strongly affected by neighbouring vowels (*cf.* the changing pattern of F2 in the voiceless [h] in Figure 3.9). From a purely acoustic point of view, the voiced [ɦ] may therefore be regarded as a vowel pronounced with a breathy voice. It is obvious, then, that the acoustic contrast with neighbouring vowels will be considerably lower, which will make segmentation more difficult. The presence of noise will affect especially formant bandwidths, which will be greater in [ɦ] than in the neighbouring vowels, as shown in the example in Figure 3.10. In this item, the “spiky” waveform associated with aperiodic noise is visible in the central portion of the fricative.

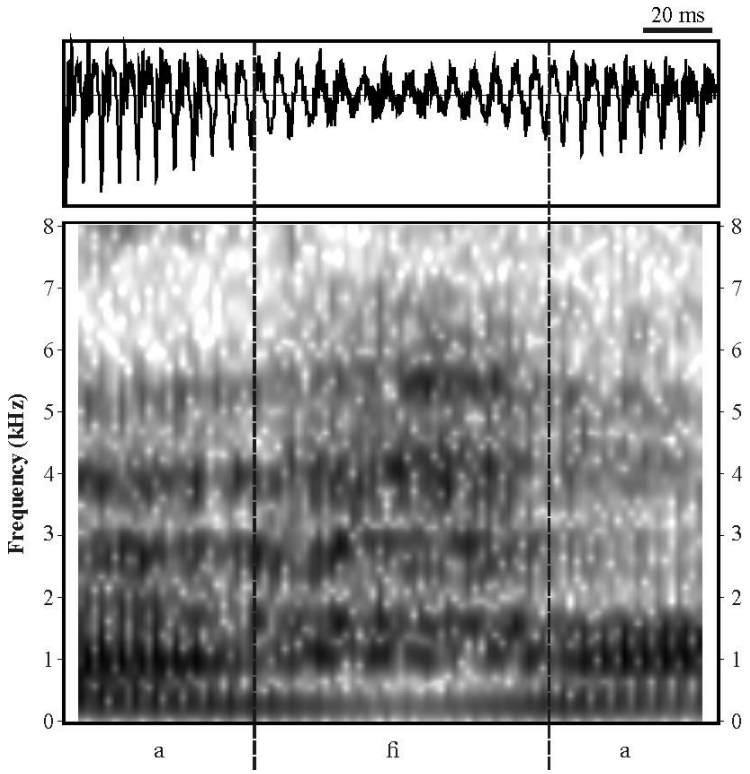


Figure 3.10. Sequence [aɦa] with salient noise and broader formant bandwidths.

As was the case with the intervocalic /v/ in Czech, lenition appears to be quite frequent in /ɦ/. In such instances, /ɦ/ loses its friction and becomes more of an approximant. From the segmentation viewpoint, the acoustic contrast between /ɦ/ and the neighbouring vowels will be lower, but some visual cues may still be available.

As for possible spectral cues to the location of the boundary, it seems that relative intensity in the region of F4 and F5 is most reliable, as indicated by the arrows in Figure 3.11. High-frequency intensity generally tends to be lower in sonorant consonants than in neighbouring vowels (*cf.* Figure 3.7 and Chapters 6 and 7).

Figure 3.12 shows an item of intervocalic /ɦ/ in which the waveform is more helpful for segmentation. The sonorant character of /ɦ/ is often associated with a simpler shape of the waveform (again, compare Chapters 6 and 7).

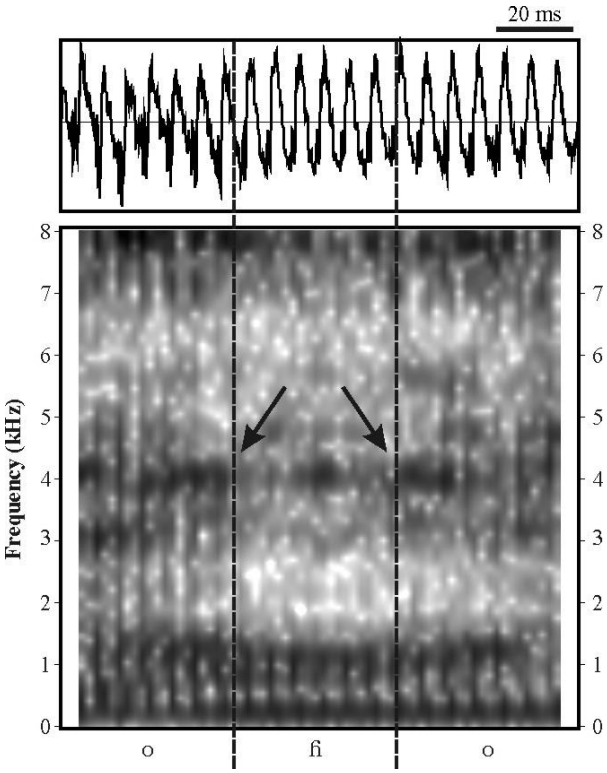


Figure 3.11. Sequence [oɦo] with relative intensity differences around 4 kHz being most helpful for segmentation (indicated by the arrows).

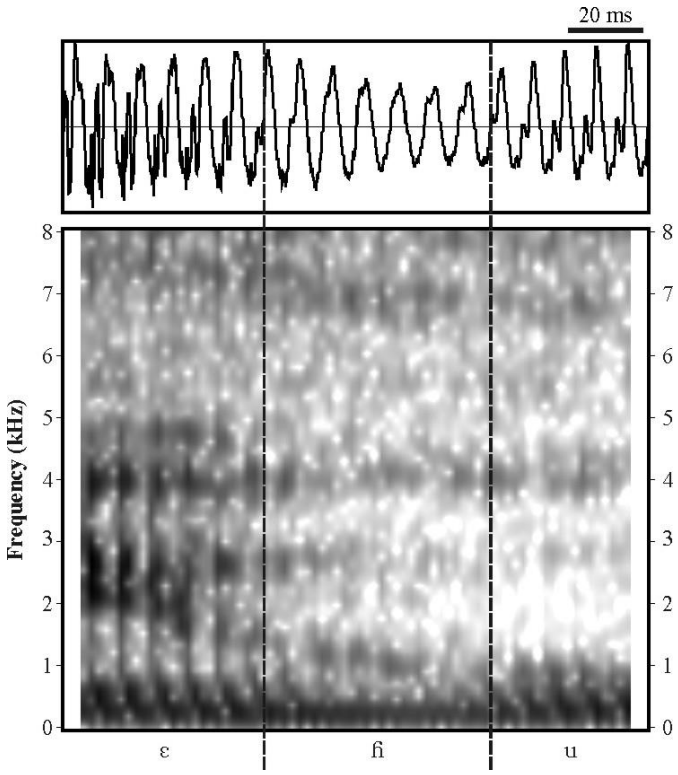


Figure 3.12. Sequence [ɛɦu] with the complexity of the waveform shape aiding segmentation.

3.5. On segmenting affricates

From the segmentation viewpoint, the left boundary of canonically realized affricates is identical to that of plosives and the right boundary to that of fricatives. The same rules apply for the particular boundaries, including cases of less explicit pronunciation. That is why no special chapter has been dedicated to affricates. Figure 3.13 shows an example of intervocalic /t͡ʃ/.

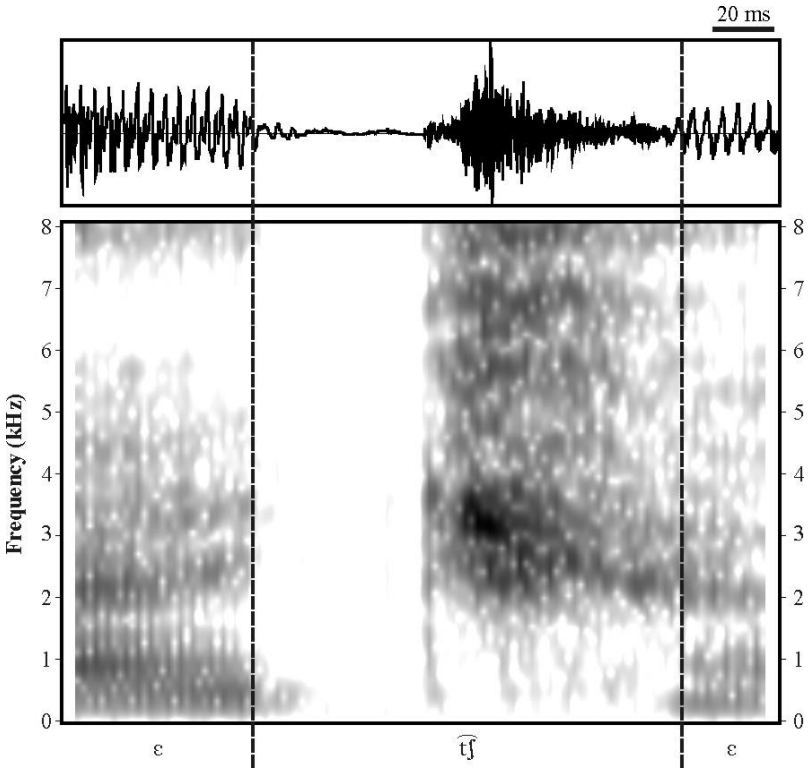


Figure 3.13. Sequence [ɛt͡ʃɛ] illustrating the segmentation of intervocalic affricates.

3.6. Summary

Aiming at highest possible consistency and comparability of segmenting various speechsounds, we consider the formant structure of the vowel to be the primary criterion in segmenting fricatives in the neighbourhood of vowels. Since both the decay and onset of formant structure may be rather slow, the boundary will often be placed near the midpoint of the transition area.

The labiodental /v/ and the voiced laryngeal /ɦ/ are often subjected to lenition in the intervocalic position. In such cases, they behave like approximants; more detail can thus be found in the relevant chapters.

4. Intervocalic nasal consonants

4.1. Articulatory and acoustic lead-in

Nasals are sonorant consonants whose articulation consists in a) the opening of the velopharyngeal port, and b) the forming of an occlusion at a specific place in the oral cavity. The former process

allows for air to pass through the nasal cavity, while the latter determines the place of articulation of the resulting nasal sound. Nasal consonants are typically voiced.

The sound inherent to all nasals is called nasal murmur, and it is described as the product of two components; naturally, these are closely related to the articulatory gestures mentioned above. The primary resonance tube is that of the pharyngo-nasal tract, stretching from the vocal folds through the pharynx, the velic opening and the nasal cavity to the nostrils. The resonance frequencies of this tube are called the nasal formants, with the average value of N1 being approximately 250 Hz and that of N2 slightly below 1000 Hz (Stevens, 1998: 489). Since the pharyngo-nasal tract does not change for a given speaker (apart from conditions such as the cold), the nasal formant values remain the same for all nasals in a given speaker. The second component of nasal spectra, the antiformant, is an antiresonance of the oral cavity which functions as a side branch. The frequency of the first antiformant, A1, differs with the place of articulation, with average values being approximately 750 Hz for the bilabial /m/, 1400 Hz for the alveolar /n/, 2500 Hz for the palatal /ɲ/, and above 3000 Hz for the velar /ŋ/. Nasal murmur is thus composed of nasal (or pharyngo-nasal) formants and oral antiformants. The presence of an antiformant is known to lower the amplitude of the spectrum above the antiformant. That is why nasal formants higher than N2 and antiformants higher than A1 are frequently not considered in acoustic descriptions.

4.2. Inherent phonetic features and basic segmentation rules

The inherent phonetic features of nasal consonants are: a) voicing and high sonority, b) a sufficient opening of the velopharyngeal port (approximately 0.2 cm^2 according to Stevens, 1998: 487; Warren *et al.*, 1993 mention the area of 0.5-1 cm^2), and c) a closure in the oral cavity (which is the reason for the presence of the antiformant).

The presence of the antiformant results in most energy being concentrated in lower frequencies. The lower intensity in higher frequencies is also caused by the greater area of soft tissue in the nasal cavity, as well as by the presence of paranasal sinuses. In intervocalic nasals, the **difference in high-frequency intensity** is, indeed, the most obvious guideline for placing the segment boundary.

The absence, or at least limited presence of higher-frequency components in the spectra of nasals also tends to be reflected in the "simpler" **shape of the waveform**. In other words, we usually cannot see the "spiky" or "hairy" character associated with high frequencies in the waveform. However, this is also the case with some vowels, for example [u].

The **antiformant** itself may show in the spectrogram as a light or even white horizontal stripe. However, antiformant valleys tend to be filled by ambient noise. Their presence can thus not be relied upon, although they can be used as a criterion for segmentation when visible.

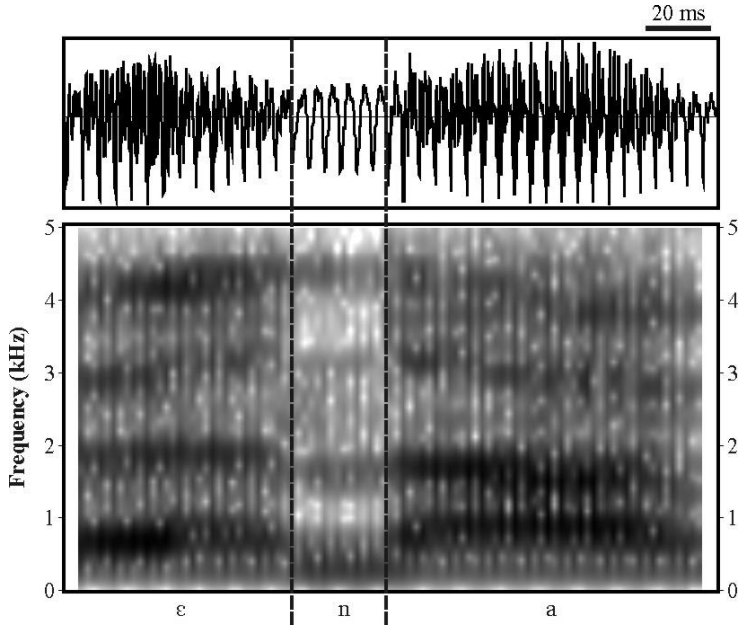


Figure 4.1. A canonical example of an intervocalic nasal in the sequence [ɛna].

Figure 4.1 shows the waveform and spectrogram of the sequence [ɛna]. It should be pointed out at the end of this section that intensity (visible in the spectrogram as darker shades of grey) may be different from peak amplitude (in the waveform). Figure 4.2 illustrates this contrast: intensity in high frequencies is obviously lower than in the surrounding vowels, while peak amplitude in the waveform is actually higher in the nasal than in most of the first vowel.

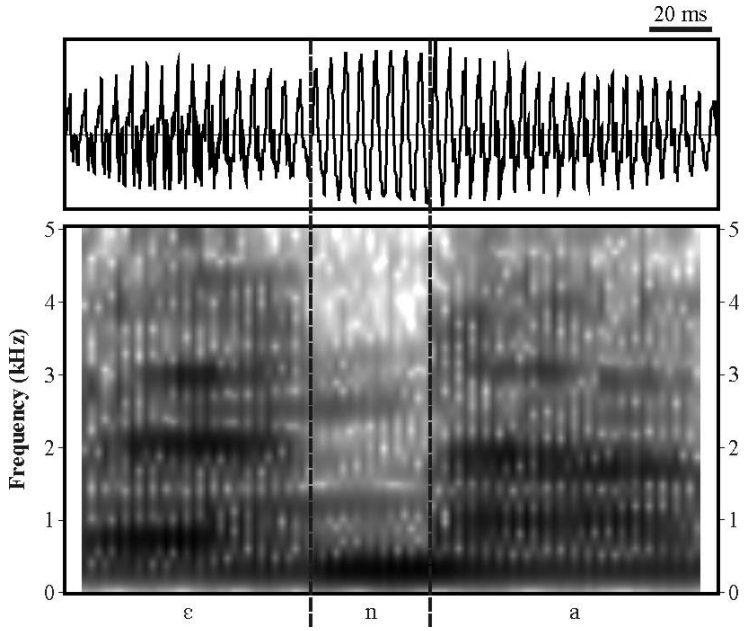


Figure 4.2. Sequence [ɛna] illustrating the apparent difference between the broad spectral intensity and peak amplitude in the waveform.

As in oral plosives (see Chapter 2), **formant transitions** may be visible in intervocalic nasals as in Figure 4.1. However, this does not seem to be frequent in fluent speech, and transitions can only rarely be exploited for segmentation.

The following two sections will present segmentation rules for the vowel-nasal and nasal-vowel boundary, respectively. As before, the boundary will be placed at the zero crossing which fulfils the criteria as much as possible.

4.3. Vowel-nasal boundary

In section 4.2, we saw that a canonical representation of a vowel-nasal boundary will be signalled by:

a) the decay of vowel formants, the decrease in spectral intensity especially in higher frequencies, and the consequent spread of formant bandwidths,

b) simplification of the waveform shape.

Moreover, it turned out that the transition from a vowel to the following nasal is often accompanied by one period with a lower amplitude. Therefore, we can postulate that c) the boundary is placed at the beginning of the lower-amplitude period, as indicated by the arrow in Figure 4.3. The beginning of this period seems to correspond best with the other boundary cues.

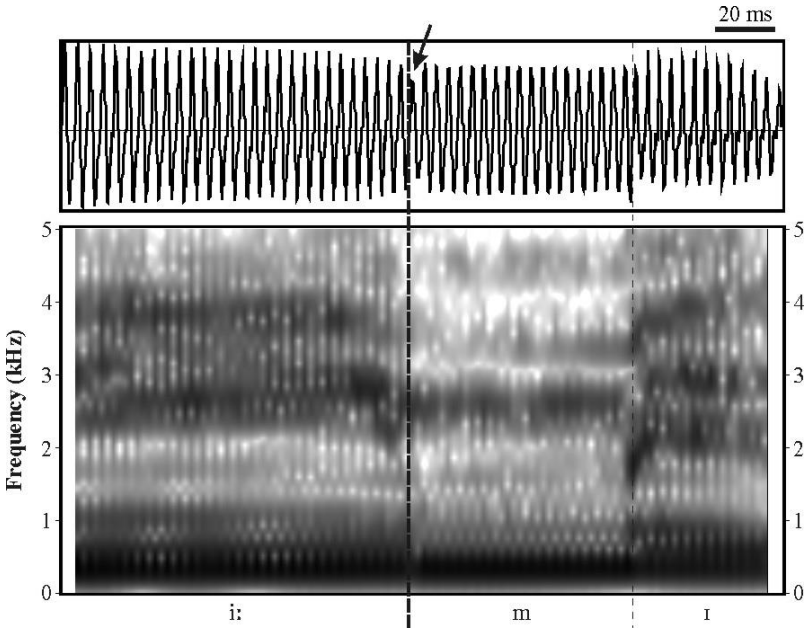


Figure 4.3. Sequence [iːmɪ] showing one period with a lower amplitude.

It may happen, and seems to happen most frequently in sequences of nasals with [iː], that no change in the shape of the waveform is visible. In such cases, we have to search for boundary cues in the spectrogram, especially the energy relationships in higher frequencies (end of vowel formants, light areas corresponding to the antiformant). Figure 4.4 shows the sequence [niːmɪ] where the boundary between [iː] and [m] (and also the one between [n] and [iː]) is not visible in the waveform.

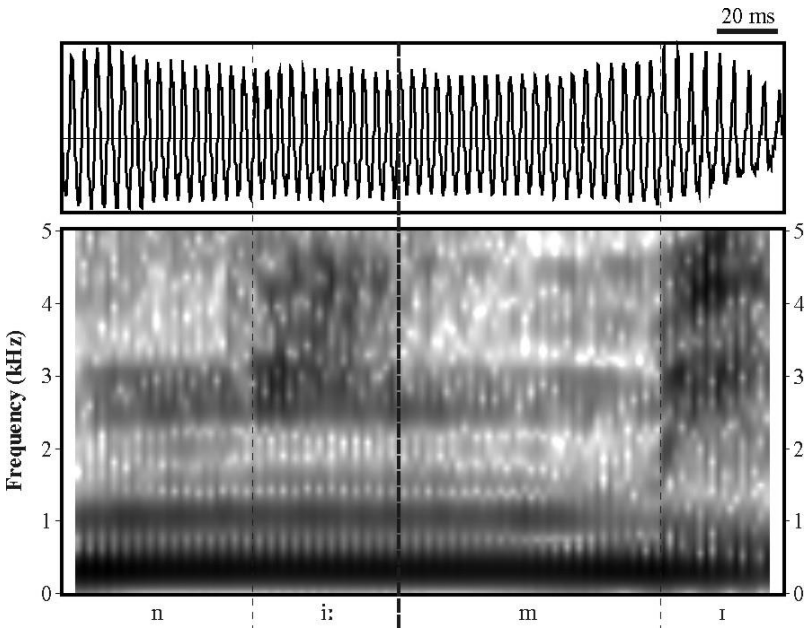


Figure 4.4. Sequence [niːmɪ] illustrating the lack of cues for segmentation in the waveform.

Naturally, in some instances we will have to exploit the rule related to the midpoint of the grey area, the transitional phase. This is shown in the interesting item in Figure 4.5.

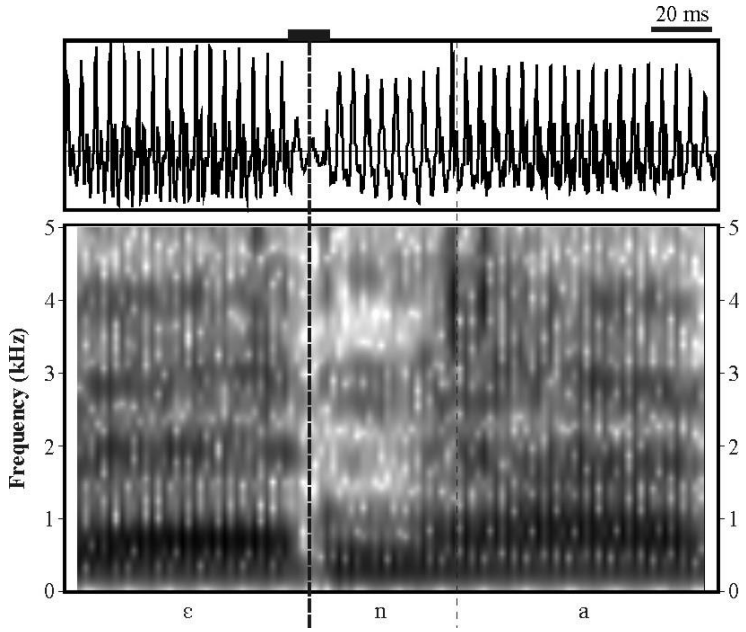


Figure 4.5. Sequence [ɛna] with the boundary between [ɛ] and [n] placed in the midpoint of the transitional phase (indicated by the horizontal black bar at the top of the figure).

4.4. Nasal-vowel boundary

Since nasal consonants may from the viewpoint of nasality be regarded as continuants, it is obvious that the transition from a nasal to a following vowel will, at least to a certain extent, display parallel cues as in vowel-nasal sequences:

a) onset of more salient vowel formants, increase in spectral intensity,

b) more complex shape of the waveform,

c) possibly the vowel transition.

Although nasals are sonorant consonants, it is sometimes possible to identify in the waveform a phenomenon resembling a plosive-like release of the occlusion. We will therefore stipulate that d) **plosion-like elements** will, in parallel with the release stage of plosives, be regarded as part of the nasal. This "plosion" in nasals seems to occur in two forms. More frequently, one can merely see one period in the waveform whose amplitude is higher than that of the preceding periods within the nasal and of the subsequent vowel (see Figure 4.6). This "plosion" usually lacks the noise component characteristic of plosions. The situation with a visible noise component appears to be considerably less common. In those cases, the plosion is also salient in the spectrogram, as shown in Figure 4.7.

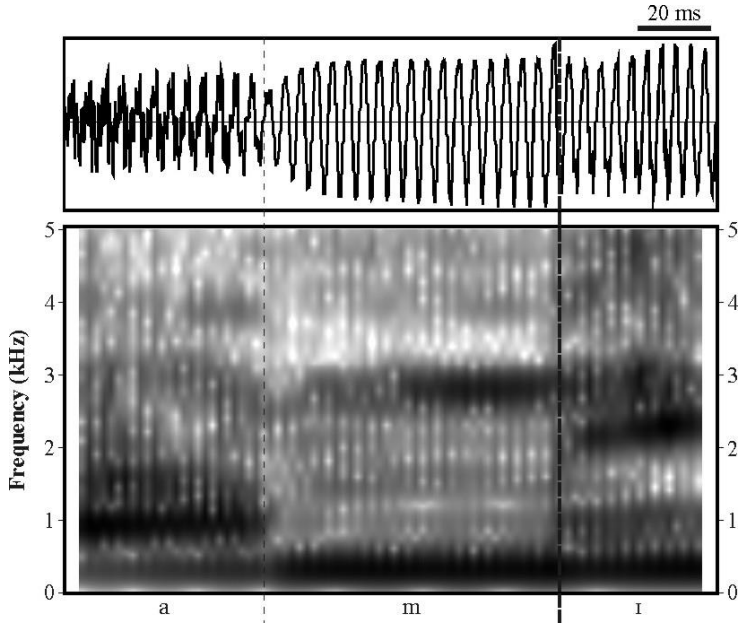


Figure 4.6. Sequence [amɪ] with a plosion-like element in the form of one period with a higher amplitude.

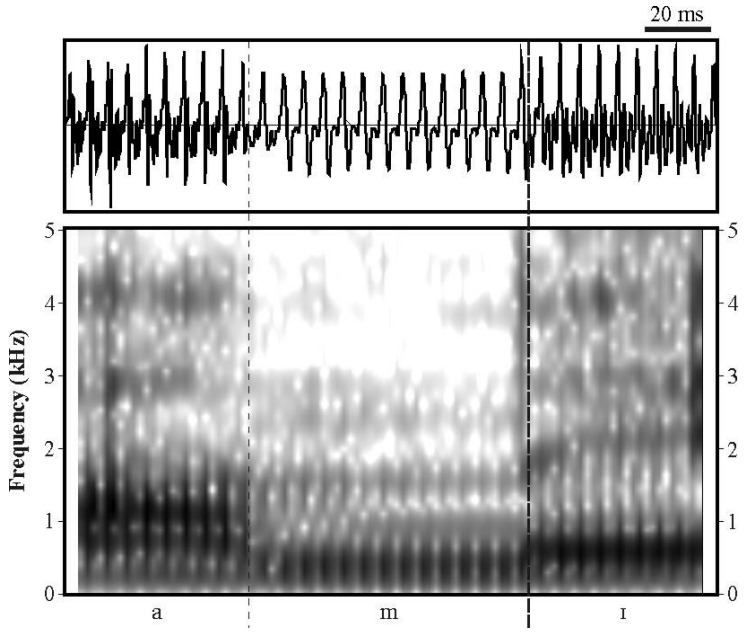


Figure 4.7. Sequence [amɪ] with a visible plosion in both the waveform and the spectrogram.

One of the most problematic sequences from the viewpoint of segmentation is that of [ɲiː]. This is caused by several reasons. First, the duration of palatal stops (both oral and nasal) is the greatest and, by the same token, so is the duration of their release phase and transitions to the next vowel). Second, as both [ɲ] and [iː] are palatal speechsounds, the changes in vowel formants are rather small. Third, the shape of the waveform tends to be almost identical. Finally, [iː] has a tendency to be heavily nasalized when adjacent to palatal sounds in general, let alone [ɲ] (Skarnitzl, 2008: 116).

Let us therefore examine the (rather few) possibilities that we have when attempting to determine the boundary location in [ɲiː]. At times, some cues can be detected in higher frequencies. As illustrated in Figure 4.8, differences in spectral intensity can be visible in the area of vocalic F4 and F5.

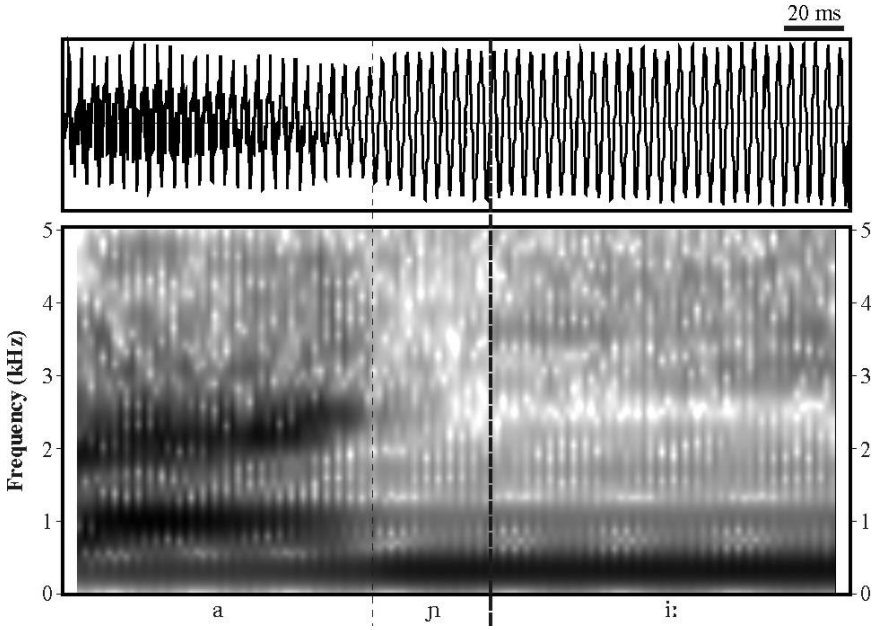


Figure 4.8. Sequence [aɲiː] with cues for boundary placement between [ɲ] and [iː] in the F4 and F5 region.

However, sometimes the acoustic contrast between the two speechsounds is so low, partly due to the gradual character of spectral changes, that reliable boundary placement, based only on visual cues, seems to be impossible. In those situations, one must also take into account listening. Figure 4.9 shows an example where we had to resort to listening, with the black bar indicating the area of gradual spectral changes.

Finally, we also have to mention instances in which no vowel has even been pronounced: for instance, the Czech phrase *jarní únava* (*spring weariness*), canonically /ˈjarɲiː ˈʔuːnava/, may be pronounced as [ˈjarɲː ˈʔuːnava]. In other words, it is impossible to detect - visually or auditorily - the release of the nasal consonant.

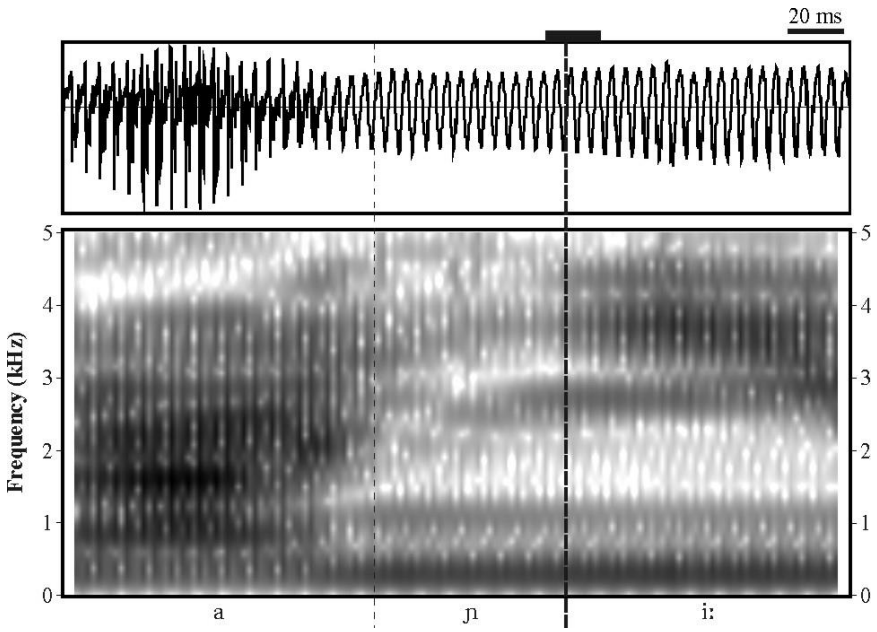


Figure 4.9. Sequence [aɲiː] with gradual changes between [ɲ] and [iː] (indicated by the black bar at the top). The boundary had to be determined with the help of auditory cues.

We have seen in Chapter 2 that the voiced alveolar plosive /d/ is often realized as a mere **flap**, [ɾ]. The same can happen with /n/ - its articulation then consists merely of a ballistic tongue movement towards the alveolar ridge, and we talk about the nasal alveolar flap, [ɾ̃]. These items are quite difficult in terms of boundary placement. There is some acoustic contrast visible in the example in Figure 4.10, but listening has to be called to aid.

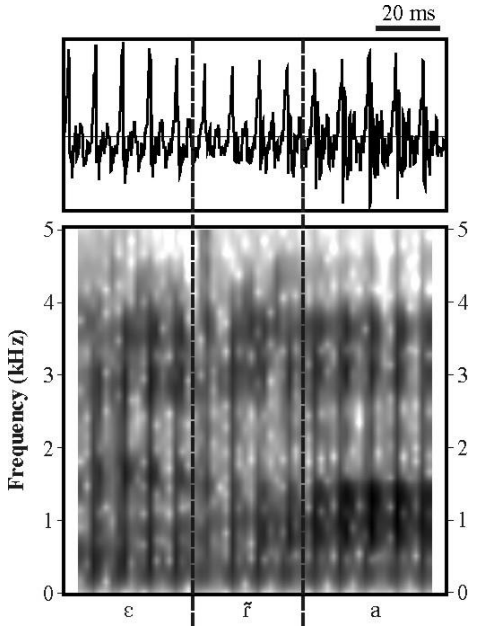


Figure 4.10. Sequence /ɛna/ in which /n/ is pronounced as a nasal alveolar flap [ɾ̃]. Listening usually has to be resorted to to aid visual cues.

4.5. Summary

To conclude this chapter, it must be emphasized that several viewpoints have to be accounted for when segmenting intervocalic nasals. Typically, we are looking for such a zero crossing which corresponds as much as possible to the changes in:

a) vowel formants (their presence in general and especially their salience) and formant transitions when they are visible,

b) spectral intensity, especially in higher frequencies (plosion-like elements are regarded as part of the nasal),

c) shape of the waveform.

In contexts where acoustic contrast is inherently low (i.e., a sequence of palatal [ɲ] with [iː] and an intervocalic nasal flap), it may be necessary to employ listening to facilitate boundary location.

5. Intervocalic trills

5.1. Articulatory and acoustic lead-in

Trills are consonants which are produced by vibrations of the active articulator with the help of the passive articulator. The active organ is appropriately positioned at the place of articulation, and the pulmonic airstream causes it to vibrate. In the languages of the world, the alveolar trill [r], in which the tongue tip is vibrating, is the most frequent, followed by the uvular trill [ʀ] and the much rarer bilabial trill [ʙ]. In the following description, we will focus only on the most typical trill, the alveolar [r], because the other trills should not differ substantially in terms of segmentation. Section 5.4 will then very briefly examine the Czech fricative trill *ř*, typically transcribed as [r̝].

The aerodynamic relationships during the production of [r] are to an extent similar to the generation of voicing. A brief approximation at the alveolar ridge is forced open by the rising pressure, allowing air to escape in a short burst. At the same time, pressure drops and the resulting suction pulls the tongue tip back to the alveolar ridge. The quasi-periodic vibration is typically described as having the frequency of 20-30 Hz. It should thus be obvious that no actual “closures”, in the sense of static events, are formed during the pronunciation of trills.

From the acoustical viewpoint, the alveolar trill [r] is neither a typical obstruent nor a typical sonorant sound. On the one hand, it is obvious that an obstacle is formed to the airstream in the vocal tract. On the other hand, [r] has a salient formant structure, with the frequency of F1 being approximately 450 Hz, that of F2 between 1300 and 1400 Hz, and F3 slightly above 2000 Hz.

During the production of a trill, we may distinguish alternating phases of relatively open vocal tract and the cycle itself (i.e., the approximation of the tongue to the alveolar ridge and its withdrawal); see section 5.2. Acoustically, these phases correspond to a vocalic part and the decay of acoustic

energy, respectively (Machač, 2009). Naturally, a trill can consist of more periods, although this is quite rare in the case of /r/ in everyday spoken Czech (see Figure 5.2 for an example of [r] with two periods).

5.2. Inherent phonetic features and basic segmentation rules

The inherent phonetic features of intervocalic trills are: a) the presence of at least one period (one approximation of the active articulator towards the passive articulator and the subsequent withdrawal, which is manifested in the signal as a sequence of vocalic part - cycle - vocalic part), and b) the presence of voicing. For the cycle itself, inherent features are marked intensity drop and weakened or absent formant structure.

Figure 5.1 shows the sequence [vr̥ɦ], in other words not an intervocalic [r], so as to better illustrate the above-mentioned phases in the acoustic signal. We can see that the phases, the vocalic parts and the cycle, are clearly demarcated in sequences with consonants. The vocalic part can assume two auditory qualities, either that of schwa [ə] (in sequences with consonants or before a pause), or that of the neighbouring vowel.

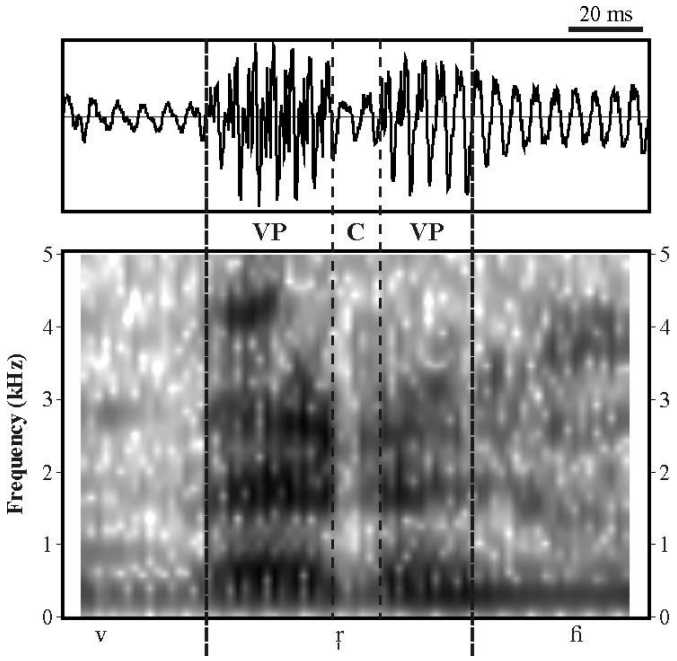


Figure 5.1. Sequence [vr̥ɦ] illustrating the phases of [r]: the vocalic parts (VP) and the low-intensity cycle itself (C).

As is obvious from the two-cycle [r] in Figure 5.2, the location of boundaries is much less straightforward in items of intervocalic [r]. While it is clear that the vocalic part is an inherent part of [r] in consonantal clusters or in pausal contexts, its “affiliation” in the neighbourhood of vowels is highly questionable. Acoustic contrast between a vocalic part of [r] and a vowel is very low and sometimes there are no visual cues whatsoever.

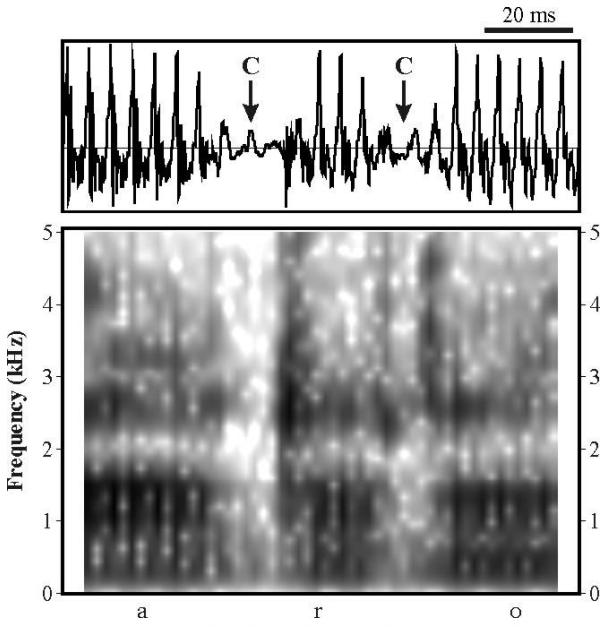


Figure 5.2. Sequence [aro] with a two-cycle [r] (the cycles are indicated by the arrows and C).

When deciding how to treat the vocalic part of [r], there is a conflict between simplicity (and thus also reliability) of segmentation on the one hand, and comparability of segmentation across different speechsound combinations on the other hand. That is why, in the following description, we will examine both possibilities.

5.2.1. The "cycle-oriented" way

Since it is quite difficult and sometimes even impossible to separate the vocalic part of [r] from a neighbouring vowel, we will consider **only the cycle itself** to constitute [r]. In other words, segmentation of [r] will differ depending on whether the neighbouring speechsound is a vowel or not. Figure 5.3 shows this more simple and reliable way of locating the boundaries of intervocalic [r]; we exploit the salient contrast between the full formant structure of the vowels and its absence in [r], along with the marked intensity drop. We can see that it would be very difficult in this example to separate the vocalic parts of [r] from the vowels on the basis of acoustic cues only.

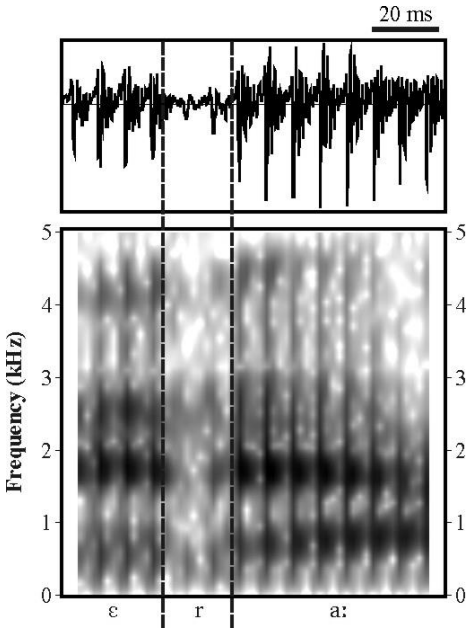


Figure 5.3. Sequence [ɛraː] illustrating the segmentation in which only the cycle constitutes [r].

5.2.2. The "extended" way

While it will always be possible to demarcate only the cycle itself as pertaining to [r], a kind of transition between the vowel and the cycle of [r] may be visible in some intervocalic items of [r]. In such items, we can regard this transition phase as **the vocalic element of [r]**.

Figure 5.4 shows such an example. We can see that the last period before the cycle and the first period after the cycle are, in both the waveform and the spectrogram, clearly different from the periods of the neighbouring vowels. In the waveform, the amplitude of those two periods is markedly lower; in the spectrogram, this is reflected in weaker intensity of the glottal pulse (the boundary between [ɛ] and [r] in Figure 5.4). One may also observe a relative difference in high-frequency intensity, in the region above approximately 2.5 kHz (the boundary between [r] and [aː] in Figure 5.4).

In other words, what can be considered to be the vocalic part may be visible in some instances of intervocalic [r]. In those cases, it seems to us better to include such periods as part of the consonant rather than the neighbouring vowel.

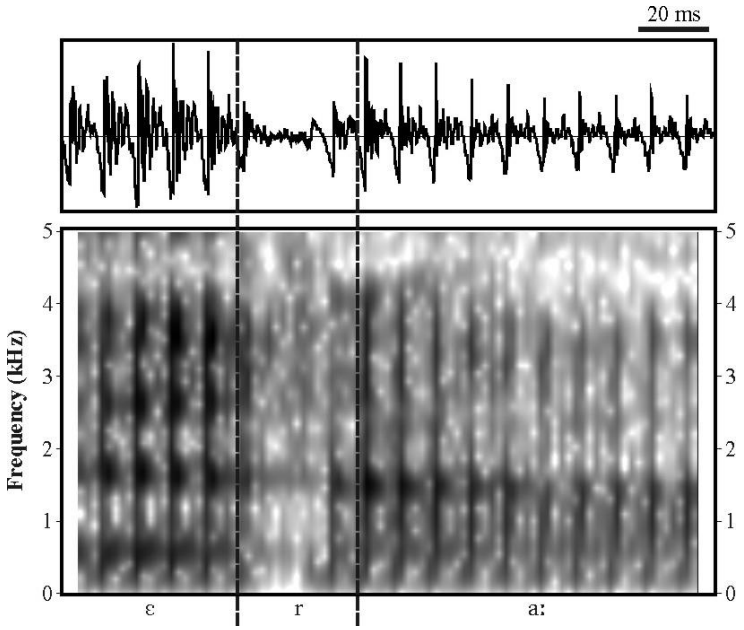


Figure 5.4. Sequence [ɛraː] with the periods surrounding the cycle different from the vocals; these may thus be regarded as the vocalic parts of [r].

5.3. Additional segmentation guidelines

Rhotic sounds in general are said to belong among the most variable speechsounds, so we are likely to encounter some extrinsic features in the signal, or the absence of intrinsic features (see page XX). First of all, the sonorant character of [r] is sometimes lost, and the cycle of [r] may even be accompanied by a **plosion-like noise**. In such instances, the right boundary of [r] will be placed at the beginning of full formant structure, as shown in Figure 5.5.

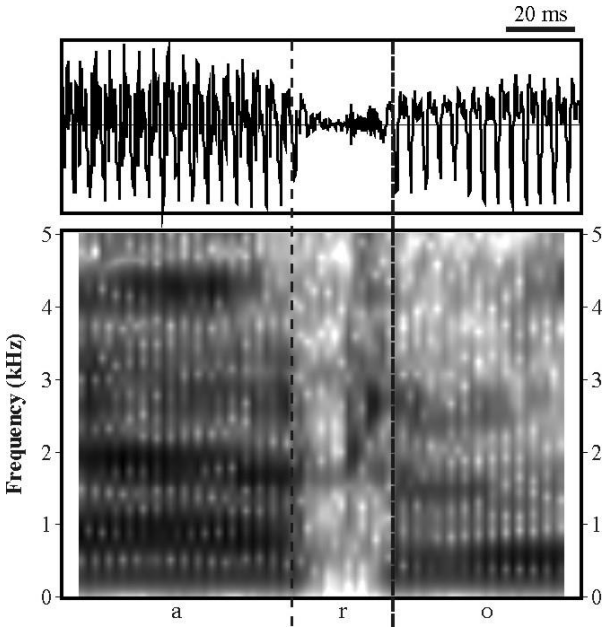


Figure 5.5. Sequence [aro] with an aperiodic noise in the cycle of [r].

It is worth mentioning that vowel-[r]-vowel sequences are often **symmetrical**, which can also be exploited for segmentation. Figure 5.6 displays also the intensity contour in which this symmetricity is visible (in spite of the aperiodic noise). The previous figures may serve as examples as well.

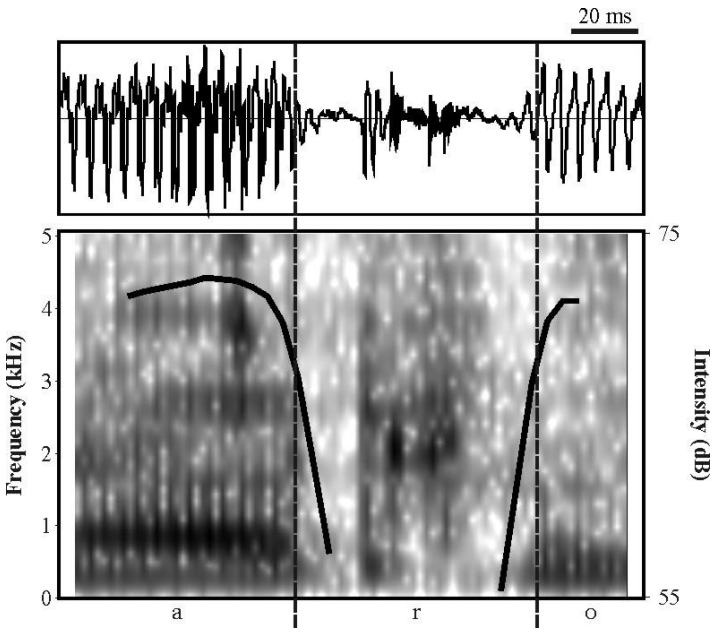


Figure 5.6. Sequence [aro] with the intensity contour indicating the symmetrical character of transitions. Symmetricity is visible both in the spectrogram (similar intensity of pulses around the boundaries) and in the waveform (in the amplitude relationships between individual periods). The thick lines representing intensity relate to the scale on the right of the figure.

In section 5.2.2, we have mentioned that one weaker period may be regarded as the vocalic element of [r]. Figure 5.7 shows that, first, the period may actually be stronger (both boundaries) and, second, there may be more periods with lower intensity in higher frequencies (right boundary). The right boundary is placed at the beginning of full formant structure, thus regarding four glottal periods as the vocalic element of [r]. It must be pointed out that this segmentation actually corresponds to the auditory impression. Figure 5.8 shows another example of more periods regarded as the vocalic element.

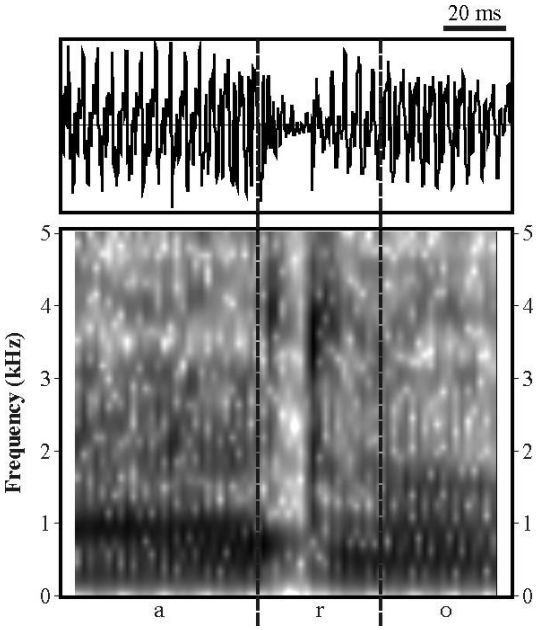


Figure 5.7. Sequence [aro] with stronger vocalic parts of [r] on both its sides, and with several periods of lower intensity in high frequencies around the right boundary.

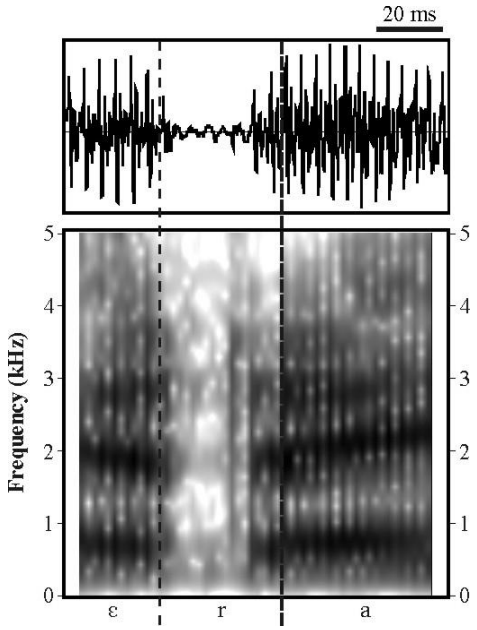


Figure 5.8. Sequence [ɛra] with a stronger vocalic part of [r] and more periods of lower intensity in high frequencies around the right boundary.

It is obvious that in implicit pronunciation, we may have to resort to the rule placing the boundary near the midpoint of the transition area. Figure 5.9 shows an item with a gradual decay and onset of formant structure, especially in higher frequencies. Apart from the formant structure itself, we can exploit the intensity contour to determine the approximate midpoint of the transition area.

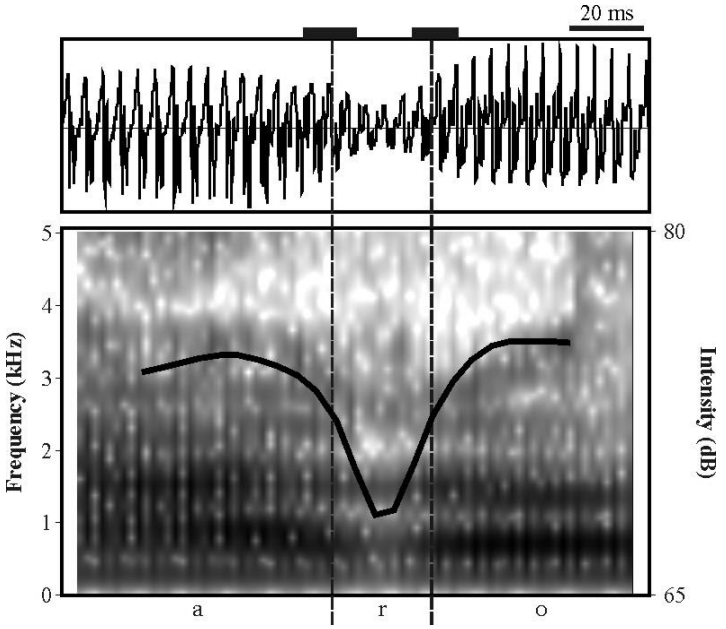


Figure 5.9. Sequence [aro] with a gradual decay and onset of formant frequencies. The boundaries of [r] are placed near the midpoints of the transition areas (indicated by black bars).

Figure 5.10 shows an item in which /r/ is realized as a uvular sound, but one of an approximant character rather than a trill. The boundaries have to be adjusted with the help of listening in this case, especially the right boundary. (Naturally, the necessity of listening sometimes applies to alveolar realizations of /r/ as well.)

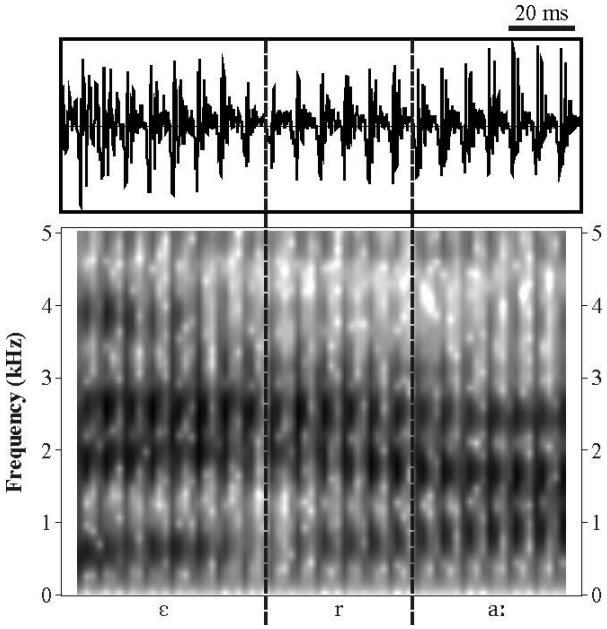


Figure 5.10. Sequence [ɛraː], or more precisely [ɛr̝aː], in which auditory cues had to be employed to determine the boundaries.

5.4. The Czech fricative trill *ř*

The fricative trill, spelled *ř*, is a sound peculiar to Czech. Today, it is transcribed with a diacritic, [r̝] (or [r̝̊] for its voiceless counterpart). It typically consists of one cycle, occasionally two cycles; the cycle is preceded by short friction and followed by longer friction. In implicit pronunciation, the cycle may be quite weak or altogether missing.

As for segmentation, we can apply the same rules for the right boundary which we apply in fricatives (see Chapter 3). In items with one cycle, the left boundary will show as a short interval of noise with the same spectral composition as the following friction of [r̝], though possibly weaker in intensity (see Figure 5.11).

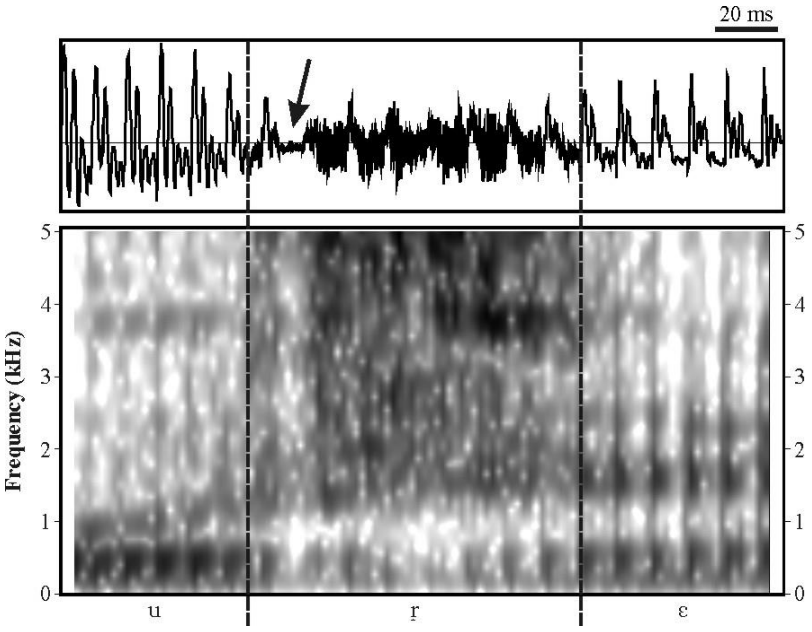


Figure 5.11. Sequence [ur̝ɛ] with one cycle of [r̝] (indicated by the arrow).

Since the short interval before the cycle itself also has a noise character, we can use fricative rules for determining this boundary as well. As with fricatives, we may have to resort to the rule placing the boundary near the midpoint of the ambiguous area.

5.5. Summary

The simplest way to segment intervocalic [r] is simply to delimit the cycle itself, i.e., the low-intensity region with suppressed or no formant structure. However, if it is possible to identify the vocalic element of [r], most frequently consisting of one period before and after the cycle, it will be phonetically more precise to separate it from the neighbouring vowel. It is advisable, in the interest of the comparability of segmentation, to indicate which method has been used.

To determine the boundaries, we will often be able to exploit the full formant structure of the neighbouring vowels rather than cues in the consonant itself. We can also exploit the frequent symmetrical character of changes in both the waveform and the spectrogram.

6. Intervocalic glides

6.1. Articulatory and acoustic lead-in

Glides are sonorant consonants, more specifically approximants. During their production, the active articulator merely approximates the passive organ, and the approximation is not as close as to result in turbulent friction.

In this chapter, we will deal with the segmentation of three speechsounds. The first two are the typical glides, or semivowels: the palatal [j] and the labio-velar [w]. The label “semivowel” refers to the fact that these two sounds are very similar in nature to the vowels of the same place of articulation, [i] and [u], respectively. Their formant values tend to be quite similar to those of [i] and [u]; if they are articulated clearly, the formants may assume even more extreme values (i.e., lower F1 for both of them, and higher F2 for [j] and lower F2 for [w]). Frequently, however, there will be no acoustic contrast between [j] and [w] and their corresponding vowels. The difference is mostly phonological, in that glides cannot function as syllabic peaks.

The third speechsound which we regard, from the segmentation viewpoint, as a glide is the English *r*-sound, the post-alveolar approximant [ɹ] of British English or the retroflex [ɻ] of American English (the difference between them is not important in terms of locating the boundaries). The most characteristic feature of these sounds is a very low F3, approximately between 1.5 and 2 kHz.

Although [j w ɹ] are all treated as glides here, we will occasionally refer to [j w] as the “true glides” in the following sections, and [ɹ] might be treated separately.

6.2. Inherent phonetic features and basic segmentation rules

The inherent phonetic features of intervocalic glides are: a) full formant structure, and b) the presence of voicing. As to the formant structure, F2 should be higher in [j] and lower in [w] than in neighbouring vowels; in [ɹ], F3 should be lower than in neighbouring vowels. In other words, there should be a slightly convex course of F2 in [j], a slightly concave course of F2 in [w], and a slightly concave course of F3 in [ɹ]. In [w], higher formants are only rarely visible.

It is obvious that the inherent features are identical to those of vowels; that is why all these sounds belong to the most problematic speechsounds from the perspective of segmentation. The spectral contrast between them and the neighbouring vowels is typically quite low, and tends to consist only in a slightly different formant pattern. Moreover, the variability of realizations is quite high, depending on articulation precision. Frequently, we will have to resort to the rule placing the boundary near the midpoint of the transition phase.

In the previous chapter, we have recommended two ways of segmenting the (canonical) intervocalic trill [r], and we will do the same for intervocalic glides. The two approaches to segmenting, which should ideally not be mixed within one study, are based on acoustic cues and perceptual cues, respectively.

6.2.1. Acoustic approach

Glides may be realized in very different ways. In explicit realizations, the slightly above-mentioned convex and concave movements of F2 and F3 should indicate the boundary. Since we are talking about speechsound combinations in which the transitions inherently take some time, the boundaries will be placed near **the midpoints of the transitions**. Figure 6.1 illustrates the application of the acoustic approach on both sides of [j] and Figure 6.2 with the right boundary of the English [w].

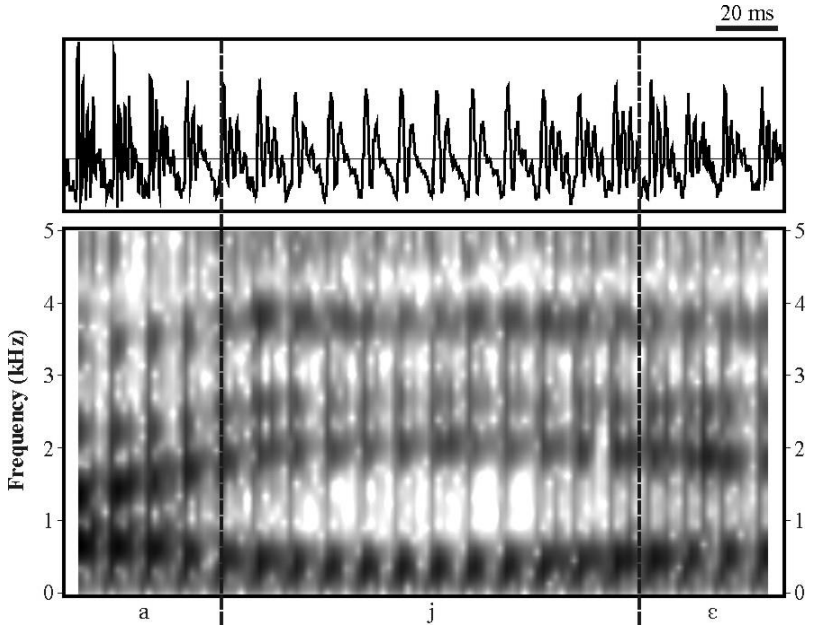


Figure 6.1. Sequence [ajɛ] in which the boundaries of [j] are placed at the midpoints of F2 transitions.

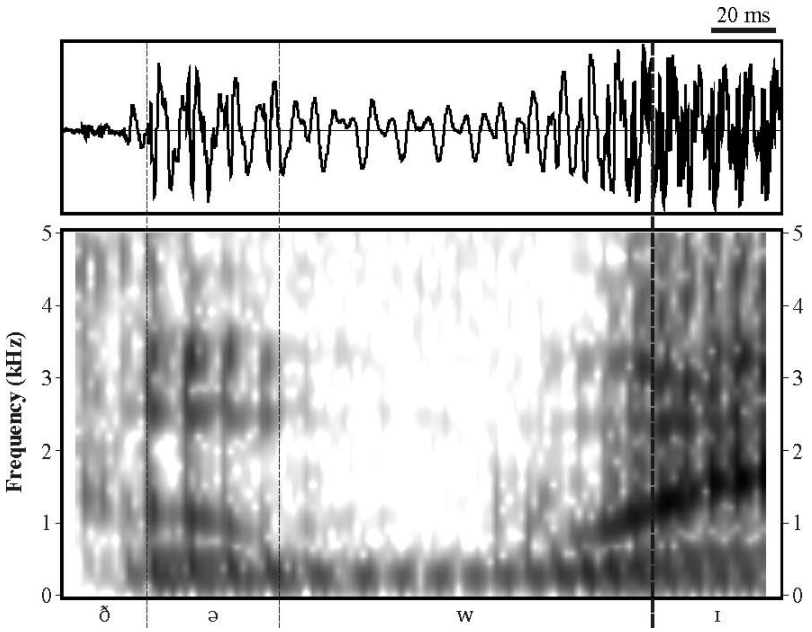


Figure 6.2. Sequence [ðə'wɪ] from the phrase *the winner* with the right boundary of [w] placed at the midpoint of the F2 transition.

The advantage of the acoustic method consists in the uniform approach to the signal, based on visual cues. Its drawback is apparent in cases when we can clearly hear a consonantal segment between the two vowels, but there are no visual cues for its identification in the acoustic signal. This will occur more frequently (but not exclusively) in sequences of the true glides with their corresponding vowel (i.e., [j] with [i] and [w] with [u]). The second drawback is the fact that when we listen to a sequence with an intervocalic glide, the acoustic method may not reflect one of the key perceptual features of glides, namely their non-syllabic character. As we will see, the acoustic approach yields glides with considerably longer duration than the perceptual approach; that is why this method of segmenting may result in the false auditory impression of syllabicity of the glide (see below).

6.2.2. Perceptual approach

In some instances, the acoustic contrast between a glide and a neighbouring vowel is so low that the auditory impression must be applied as the primary guideline, with visual information regarded merely as auxiliary. Let us examine, in more detail, one such example of intervocalic [j] in Figure 6.3. Although the spectrogram does not contain reliable cues of the presence of [j], it is quite salient auditorily.

When locating the boundary by means of listening, the task is to find the moment when we still can hear the sequence /oj/ or /ja/ as monosyllabic (and not as a sequence of two syllables). When we want to locate the right boundary of [j], we try placing the boundary further to the right, into [a]. Then we start shifting the boundary in the transition phase between [j] and [a] leftwards, according to the auditory impression, until we can hear a monosyllabic (diphthongal) sequence [oj], not something like [oj^{ə}] (i.e., no vocalic element). The left boundary will be located analogously: we place the boundary into [o] and proceed to the right, until we hear monosyllabic [ja] and not a disyllabic [^{ə}ja]. The application of this "careful listening" procedure leads to the segmentation shown in Figure 6.4. Obviously, we can still hear transitions of [j], especially in the following vowel.

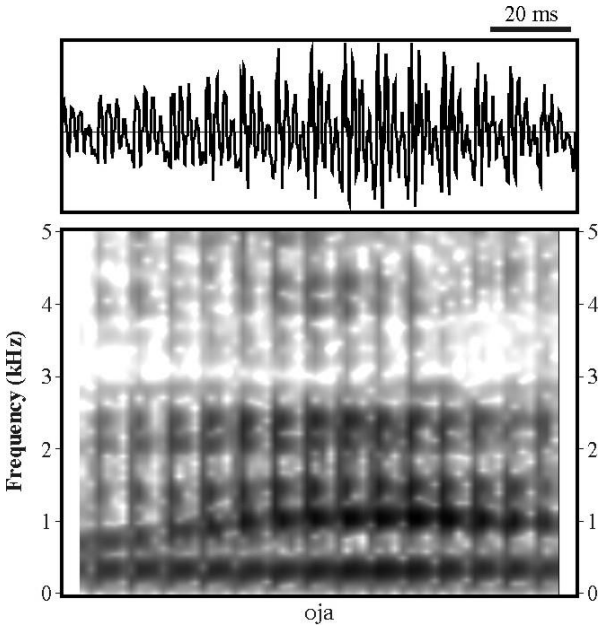


Figure 6.3. Sequence [oja] illustrating the lack of reliable cues for segmentation.

The advantage of the perceptual method is its universal character, in that it uniformly applies not only to straightforward cases, but also to unclear cases in which we can hear [j] or "something like [j]" although there are no obvious visual cues for its segmentation available in the spectrogram. On the other hand, this approach is time-consuming, demanding in terms of the labeller's concentration and, naturally, more subjective.

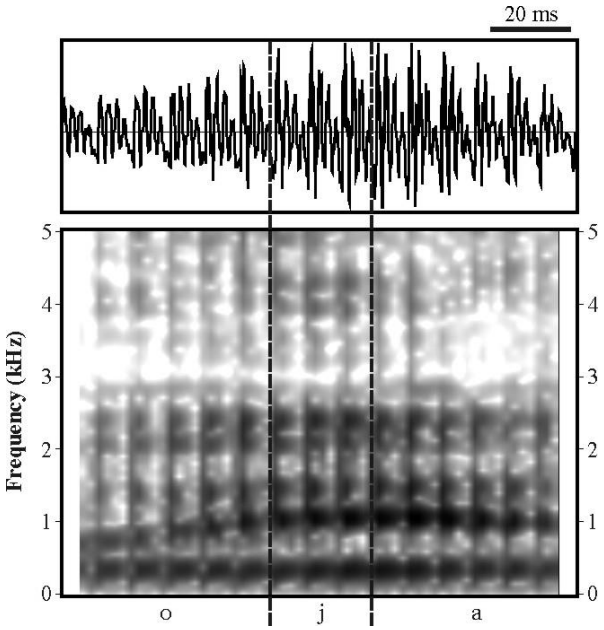


Figure 6.4. Sequence [oja] illustrating the application of the perceptual approach.

6.3. Additional segmentation guidelines

It appears, at least from the material we had at our disposal, that the English [w] and [ɹ] do not present the labeller with too many difficulties. There seem to be sufficient cues for segmentation even in cases of relatively implicit pronunciation. These include not only formant movements (F2 for [w] and F3 for [ɹ]), but also formant intensity. Figure 6.5 shows an example of [ɹ] where the movement and intensity of F2 and F3, as well as intensity in the F4 and F5 region, allow for relatively reliable boundary placement. (The acoustic approach will be applied in this section, unless specified otherwise.)

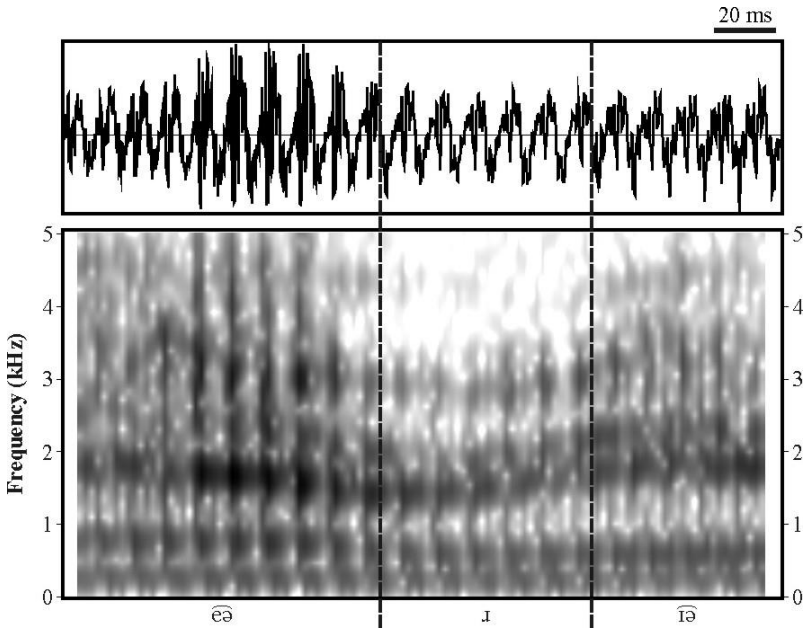


Figure 6.5. Sequence [ˈɛə͡ɹɪə͡] with formant movement and intensity providing segmentation cues.

In the labio-velar [w], the labialization is sometimes so intensive that higher frequencies are basically invisible. This is apparent to a lesser degree in Figure 6.6 and quite clearly in Figure 6.7, where the prevalence of low-frequency energy is obvious not only from the spectrogram, but also from the smooth course of the waveform.

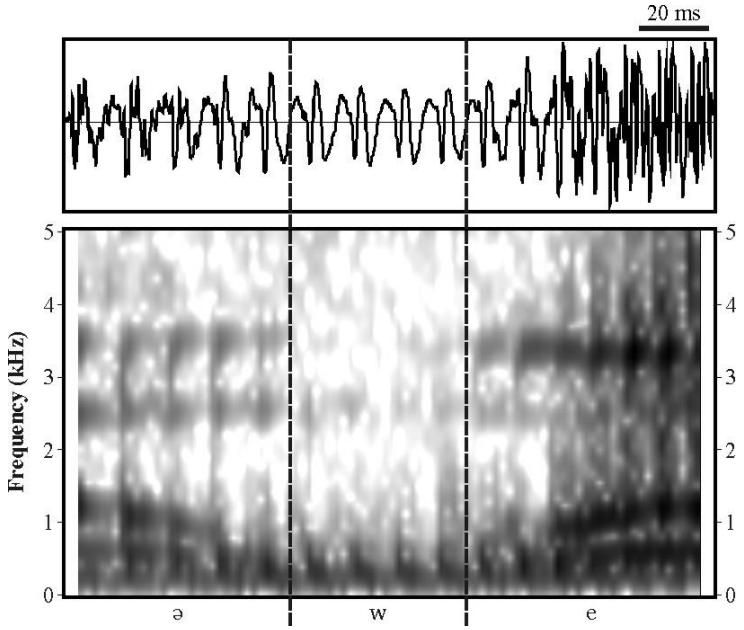


Figure 6.6. Sequence [əˈwe] from the phrase *a wealthy* with most of the energy present in low frequencies.

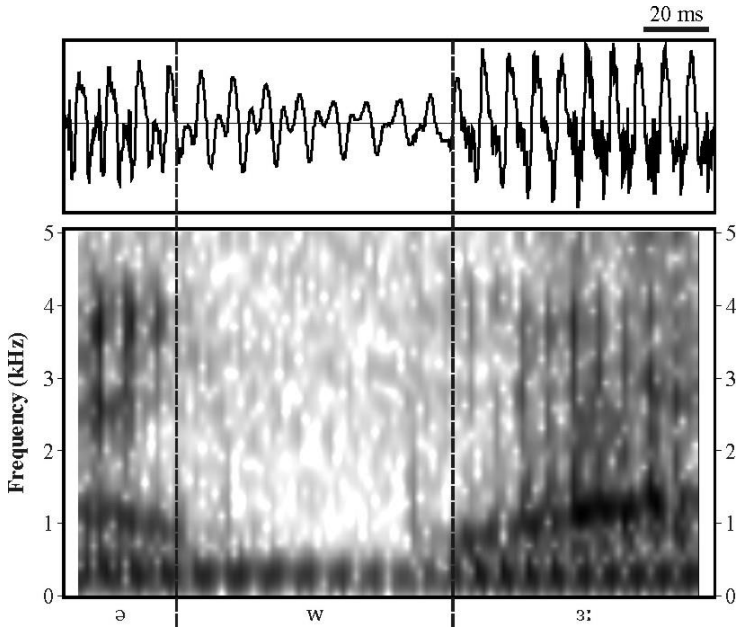


Figure 6.7. Sequence [əˈwɜː] from *the world* with most of the energy present in low frequencies.

It seems that even in sequences of [w] with its "twin" vowels, [u] or [ʊ], boundary location is not too challenging. Figure 6.8 shows an item with quite fast pronunciation in which [w] appears, unlike in the previous examples, at the beginning of an unstressed syllable. In spite of this position of low prominence and fast realization, segmentation is not problematic according to the criteria mentioned in section 6.2.

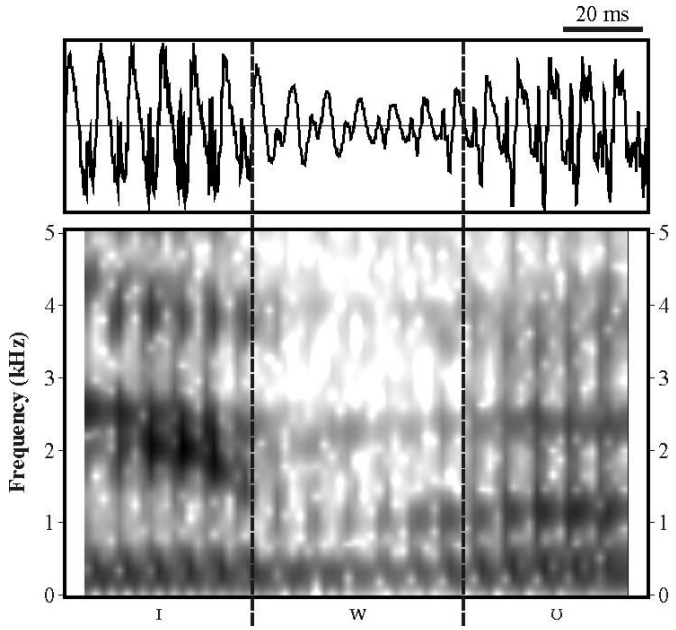


Figure 6.8. Sequence [ɪwʊ] from the sequence *he would*.

Obviously, we may encounter items in which listening will be required when segmenting intervocalic [w] or [ɹ]. Figure 6.9 shows an example with [ɹ] in which the signal does not provide any reliable cues; the perceptual approach had to be used in this case.

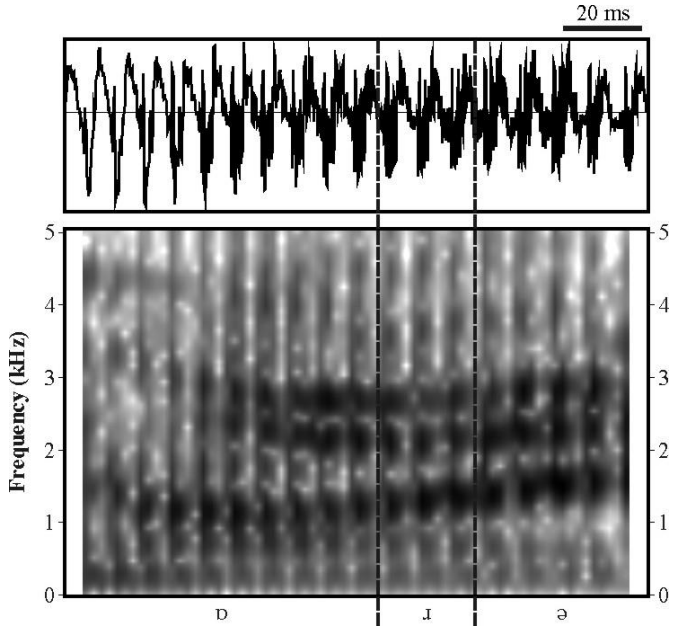


Figure 6.9. Sequence [ɒɹə] from the word *correspondent*; the perceptual approach had to be used to locate the boundaries of [ɹ].

Our material suggests that /j/ in Czech is more complicated from the segmentation viewpoint than the English /w/ and /ɹ/ (or the English /j/, for that matter). This may be partly caused by the different contexts in which /j/ occurs in Czech: it is very frequent in unstressed syllables, in bound morphemes, and thus seems to be more "vulnerable". In the rest of this section, we will therefore deal with locating the boundaries of less explicit (or otherwise noteworthy) items of the Czech /j/.

Realizations of /j/ vary from very explicit, careful pronunciation to implicit cases where one can not even be sure whether a consonant has actually been pronounced. First, we will look at the former case. Figure 6.10 shows what we may call “fortis” pronunciation of /j/ in which the approximation of the tongue to the palate is so intensive that it results in audible friction. Phonetically, we can talk about a voiced palatal fricative, [j̬]. For locating the boundary, this explicit realization is, in fact, easier than the gliding version: we follow the criterion of full formant structure, as with fricatives. In [j̬], formant structure is weaker and spectral intensity in general is lower.

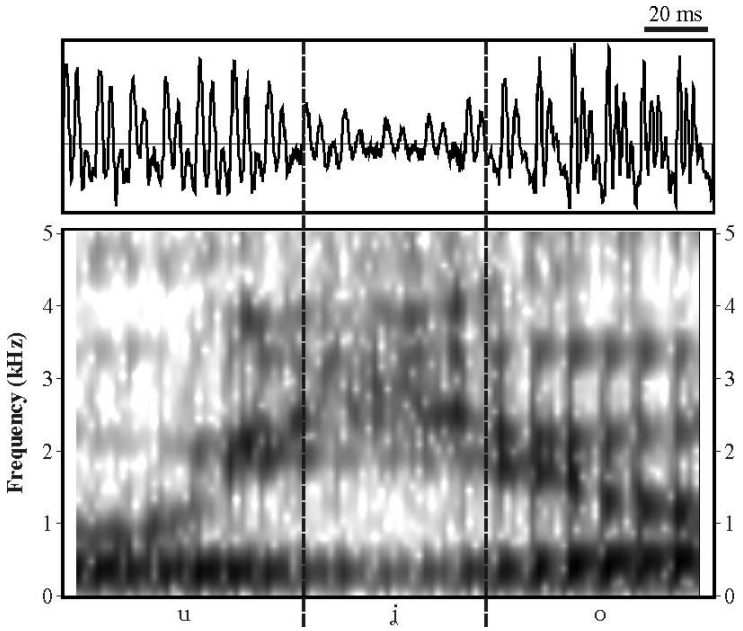


Figure 6.10. Sequence [uj̬o] with a fricative-like /j/.

A relatively explicit, but not fricative realization of /j/ may be manifested in the signal by lower amplitude. This appears to be frequent especially when the neighbouring vowels are both high and front (/ɪ iː/). Figure 6.11 shows an example of this, with the boundaries placed near the midpoint of the amplitude envelope changes (as well as the F2-intensity changes).

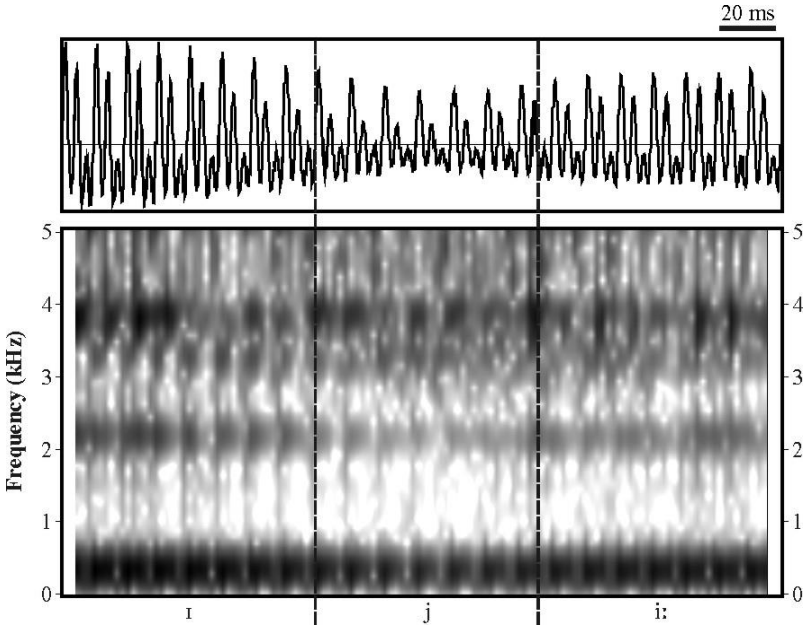


Figure 6.11. Sequence [ɪjiː] with changes in the amplitude envelope and F2 intensity facilitating segmentation.

Naturally, less explicit realizations are much more frequent. The tongue body often does not achieve the target articulatory position. In such cases, the perceptual approach outlined above seems to be more helpful than the acoustic approach. Figure 6.12 shows another instance of the sequence [ɪjiː] in which it is impossible to determine boundary placement without listening.

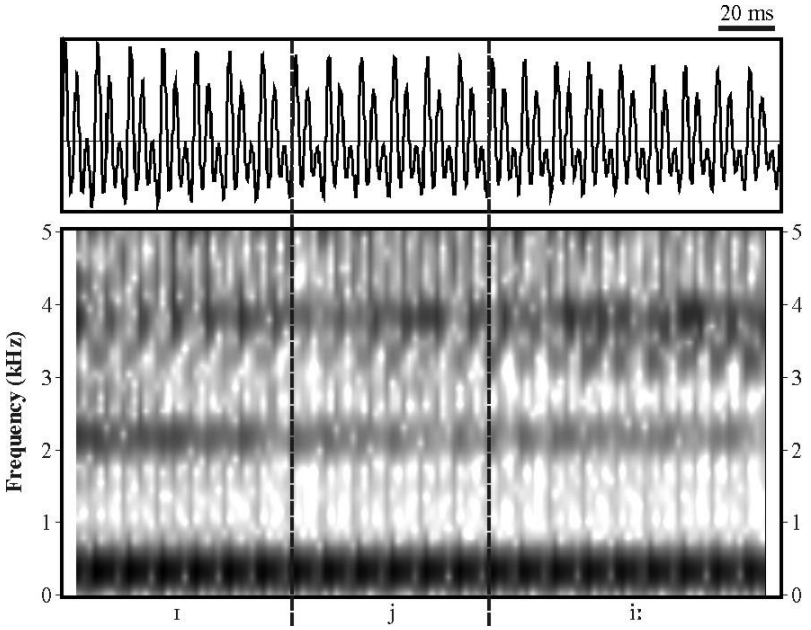


Figure 6.12. Sequence [ɪjiː] in which auditory cues had to be exploited to locate the boundaries of [j].

6.4. Summary

To segment intervocalic glides, we may choose between the acoustic or perceptual approach. The choice will probably depend on our research objectives and, perhaps more importantly, on the nature of our sound material: it appears that glides in English are less prone to such target undershoot as to completely rule out the acoustic approach.

In the acoustic approach, we place the boundaries near the midpoint of the transition between the glide and the neighbouring vowel. This transition concerns F2 in the case of [j] and [w], and F3 in the case of the English approximant [ɹ]. The midpoint of amplitude envelope changes may also be exploited in the acoustic approach.

In the perceptual approach, we segment the signal so that the sequences vowel-glide or glide-vowel sound as one syllable, as a diphthong.

7. Intervocalic lateral alveolar approximant

7.1. Articulatory and acoustic lead-in

Laterals are typically sonorant consonants which are articulated with a closure along the mid-sagittal line of the oral cavity with the sides of the tongue lowered, so that air can flow along them (or along one of them in the case of the so-called unilateral laterals). The most frequent lateral sound is a voiced lateral alveolar liquid; that is why it will be the focus of this chapter.

Lateral approximants are quite challenging from the viewpoint of segmentation because, first, they frequently undergo coarticulation so that their acoustic qualities "seep" into neighbouring sounds and, second, they are often characterized by high acoustic variability. With /l/, for example, Czech has seen lately an increase in the pronunciation of velarized [ɫ] (see, e.g., Volín, 2002). In English, velarization is systematic in other than prevocalic positions. In extreme cases, the alveolar contact can be lost, leading to the vocalization of /l/. On the other hand, /l/ in German is a palatalized sound [l^{j}]. As we will see, these sources of variability mean that it is virtually impossible to specify a general set of hierarchical rules for the segmentation of intervocalic /l/.

As the lateral approximant is a sonorant sound, its spectrum will be characterized by a formant pattern. Approximate values of the first three formants (for the canonical alveolar /l/) are 350, 1300, and 2800 Hz, respectively (Stevens, 1998: 546). The bifurcation of the air stream in the oral cavity means that an antiformant should be visible in the spectrum of /l/. The antiformant tends to appear between 2 and 3 kHz, depending on the length of the side branch.

7.2. Inherent phonetic features and basic segmentation rules

The relevant inherent phonetic features of the alveolar lateral approximant are: a) voicing and high sonority, b) a closure in the alveolar region, and c) lowering of the tongue sides resulting in the bifurcation of the air passage (which is the reason for the presence of the antiformant).

In /l/, the inherent features by themselves allow us to stipulate segmentation guidelines which will be essentially identical to those for nasals (*cf.* section 4.2). The **antiformant** appears to be salient in the 2-3 kHz region in a considerable part of instances of intervocalic /l/. The presence of the antiformant should cause a **drop in high-frequency intensity** (approximately in the F4 and F5 region) and a "simpler" **shape of the waveform**.

Unfortunately, it seems to be rather rare for these three cues to manifest at the same time. Figure 7.1 shows the waveform and spectrogram of one such example taken from the Czech name *Alena*. Since waveform cues are not available very often in items of intervocalic /l/, figures in this chapter will always present the spectrogram, but the waveform will be displayed only when it can be exploited for segmentation purposes.

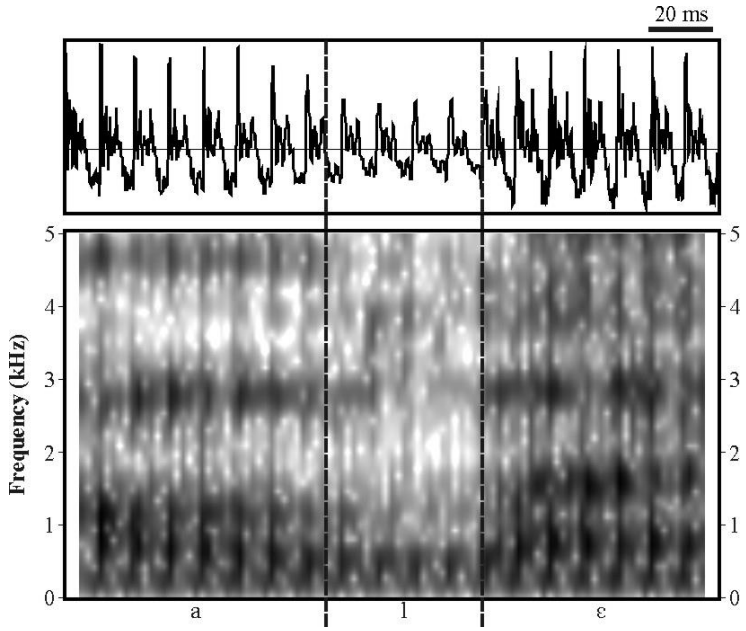


Figure 7.1. A canonical example of an intervocalic /l/ in the sequence [alɛ] with all three primary cues available for segmentation (see text).

In the following section, we will look at more typical (and less canonical) instances of /l/ and try to identify the cues that may be exploited for locating the boundaries.

7.3. Other segmentation guidelines

We have mentioned in the previous section that the three inherent cues are rarely all visible at once. Figure 7.2 illustrates an example in which the antiformant is present and high-frequency intensity is also lower in the consonant, but the waveform does not provide any cues. Figure 7.3a shows an antiformant (albeit relatively weak) but no intensity contrasts in high frequencies, while Figure 7.3b presents the opposite case with a high-frequency intensity contrast but without a visible antiformant.

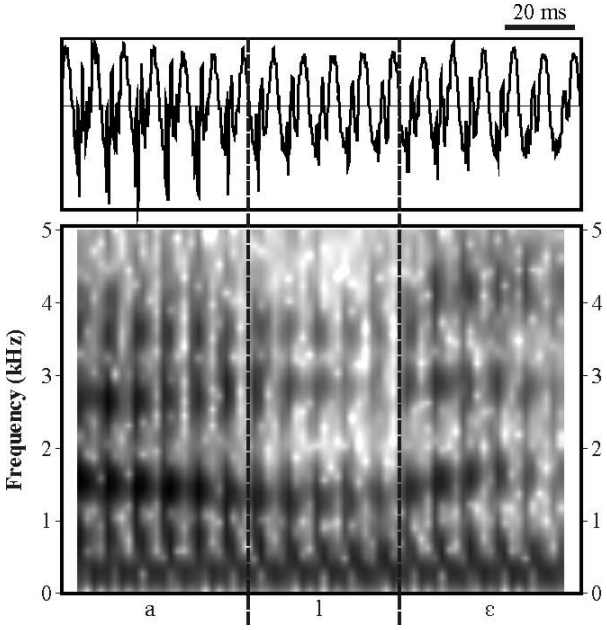


Figure 7.2. Sequence [alɛ] with no cues available in the waveform.

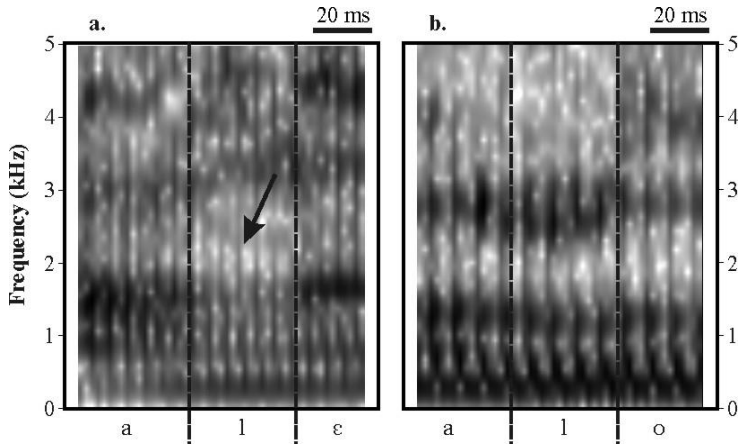


Figure 7.3. a. Sequence [alɛ] with a visible antiformant (indicated by the arrow) but no intensity contrast in high frequencies; **b.** sequence [alo] with a high-frequency contrast but no clearly defined antiformant.

It appears, however, that the cues derived on the basis of the inherent phonetic features are not the most reliable ones from the viewpoint of segmentation. We have shown (Skarnitzl, 2009) that **relative intensity of formants** can be exploited most frequently in the intervocalic /l/. Given the presence of the antiformant, lower intensity of formants is to be expected, especially in higher frequencies. It was surprising to find the exact opposite with F3: in many instances, F3 turned out to be more salient in /l/ than in the neighbouring vowels (Figure 7.4). Apart from F3, relative formant intensity can also concern F2 - F2 is typically weaker than in neighbouring vowels (Figure 7.5).

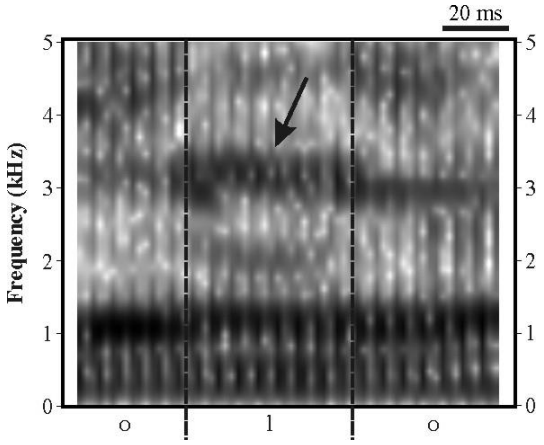


Figure 7.4. Sequence [olo] with F3 more salient (indicated by black arrow) in /l/ than in the neighbouring vowels.

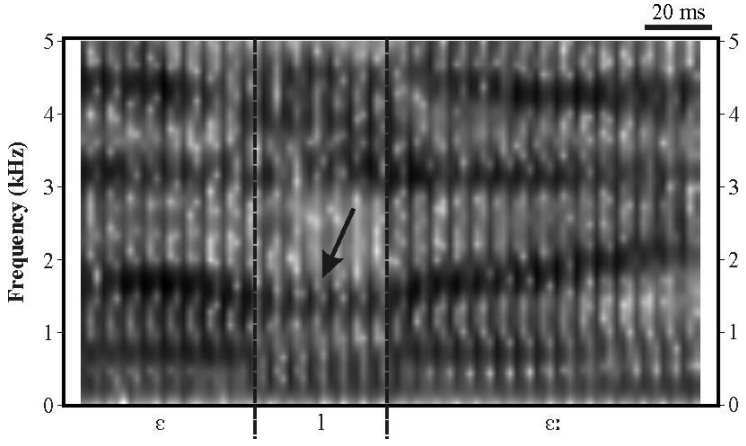


Figure 7.5. Sequence [ɛlɛː] with F2 weaker (indicated by black arrow) in /l/.

The shape of the waveform as a possible cue for segmentation has been mentioned above. Figure 7.6 shows a sequence in which the waveform is actually more helpful than spectral cues (especially in the case of the left boundary). The shape of the periods in /l/ is quite distinct, and also peak amplitude is higher than in the neighbouring vowels.

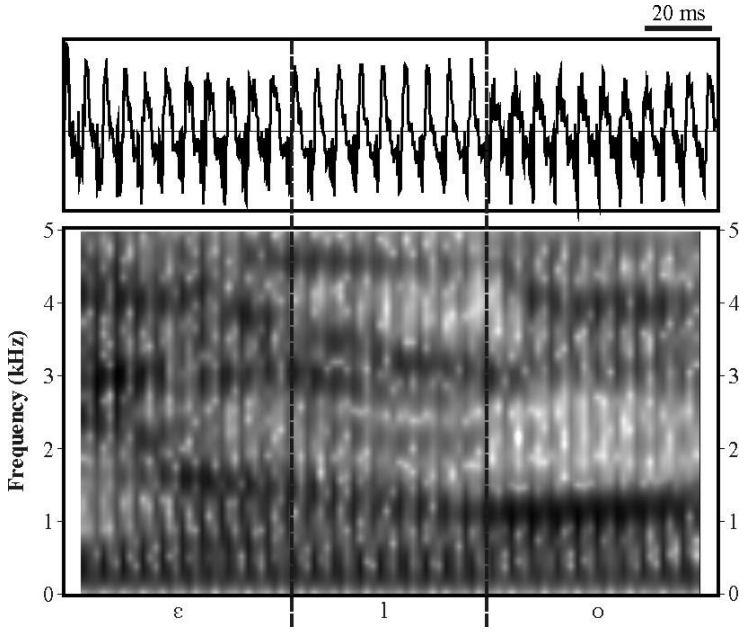


Figure 7.6. Sequence [ɛlo] with more cues for segmentation available in the waveform than in the spectrogram.

In some situations, especially when the vowel following /l/ is a high front vowel [ɪ iː], the release of the alveolar contact may be visible in the spectrogram as a **plosion-like element** (*cf.* the similar phenomenon in Figure 4.7 for nasals). This is illustrated in Figure 7.7, which shows both the waveform and the spectrogram of the sequence [alɪ].

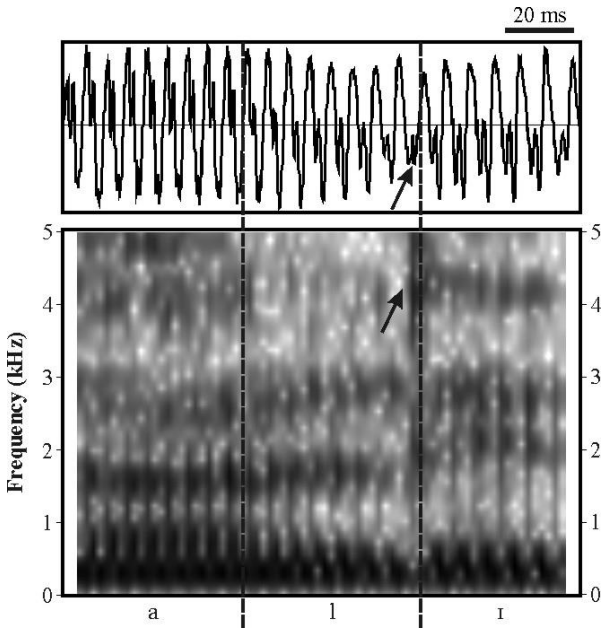


Figure 7.7. Sequence [alɪ] with a plosion-like element (indicated by the black arrows).

We have already mentioned that /l/ is one of the most problematic sounds for segmentation. It is not surprising, then, that a particular item may require the attention to auditory cues. This happens especially in situations where /l/ is vocalized, and acoustic contrast with neighbouring vowels is therefore low. Figure 7.8 illustrates one such example in which spectral cues for segmentation of the right boundary are quite weak, and we have to complement them with listening. Figure 7.9 shows the sequence /ɛlela/ in which auditory cues are indispensable for segmentation.

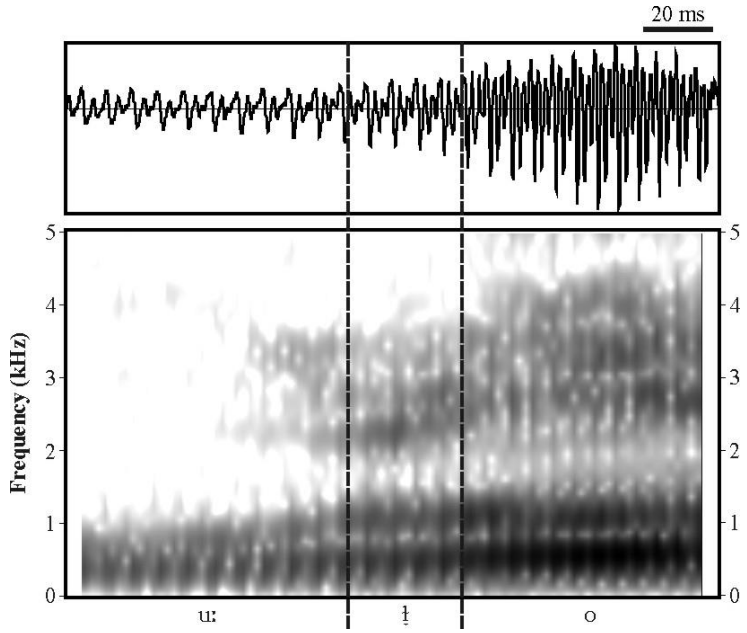


Figure 7.8. Sequence [uːɬ̪o] where visual location of the right boundary had to be aided by listening.

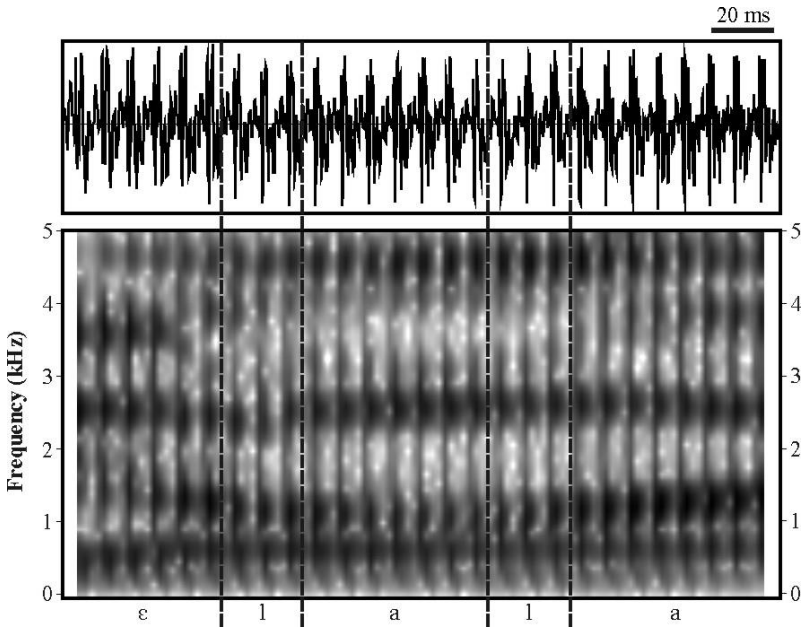


Figure 7.9. Sequence /ɛlala/ where listening is more important for locating the boundary than visual information.

7.4. Summary

It should be obvious from the preceding description that individual acoustic cues may combine in different ways. One frequently encounters items in which the left boundary of /l/ is easily detectable, for instance, from the shape of the waveform, while the right boundary will be clearly marked by intensity differences in the F4 and F5 region.

As we have hinted at in the introduction to this chapter, a set of hierarchical segmentation rules is not feasible for /l/. We can, however, say how reliable the above-mentioned cues appear to be, in other words, how often they are available for segmentation (based on Skarnitzl, 2009):

a) relative formant intensity (typically weaker F2 and often stronger F3 in /l/),
b) intensity contrast in high frequencies (F4 and F5 region),
c) simpler shape of the waveform,
d) clear presence of the antiformant,
e) visible release of the alveolar contact.

8. Obstruent clusters of different manner of articulation

8.1. Articulatory and acoustic lead-in

This chapter is dedicated to consonantal clusters of fricatives and plosives, possibly also affricates. Fricatives are characterized mainly by the presence of noise whose spectral composition depends predominantly on the place of articulation. The closure phase of voiceless plosives corresponds to a "silent" portion of the signal, while in voiced plosives a periodic component is visible with no discernible formant structure. The release phase of voiceless plosives typically features a noise burst; the noise component tends to be less salient, or may be completely absent, in voiced plosives. For more detail on these consonant classes, see Chapters 2 and 3.

As in Chapter 3, the illustrations will display the 0-8 kHz frequency range, so that a greater contrast is visible, especially in alveolar fricatives.

8.2. Basic segmentation rules

To separate fricatives or affricates from plosives, we regard the presence of **full fricative noise** as the primary phonetic feature. Illustrations may be found in Figures 8.1 to 8.4 for various combinations of these sounds, also with regard to voicing.

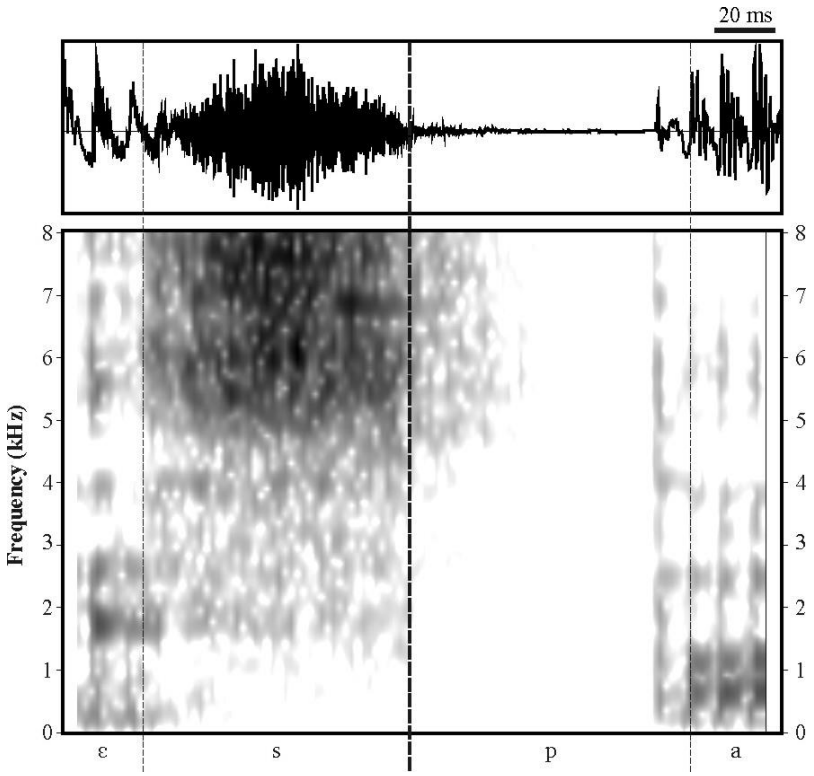


Figure 8.1. Sequence [ɛspa] with the boundary between [sp] clearly indicated by the noise.

This approach appears to be reliable even when segmenting implicit pronunciation, since the presence of the noise component is quite stable, especially in alveolar and post-alveolar fricatives (Machač, 2004), even in strong deformations of speech like spirantization of plosives (see Figure 8.9 for an example).

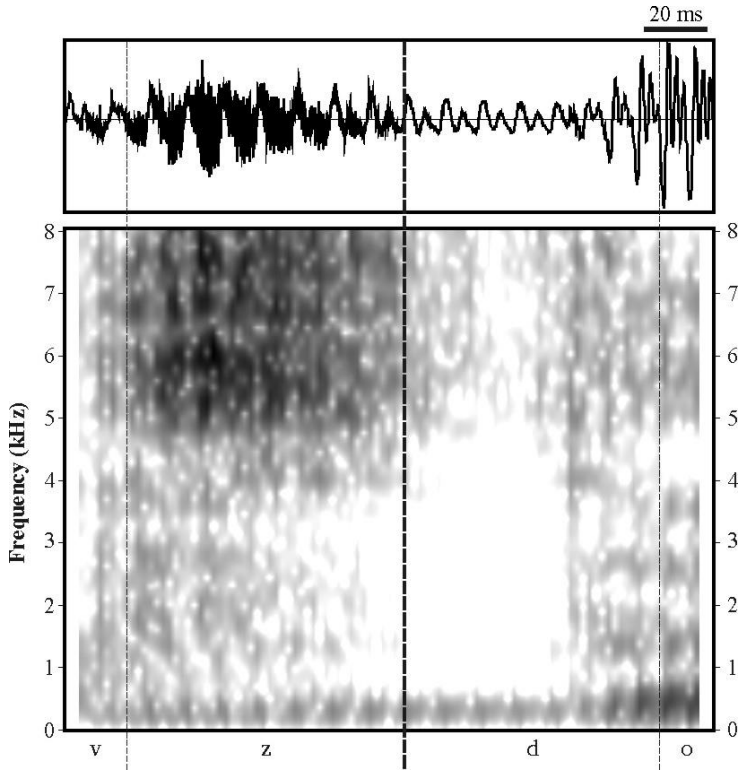


Figure 8.2. Sequence [vzdo] with the boundary between [zd] clearly indicated by the noise.

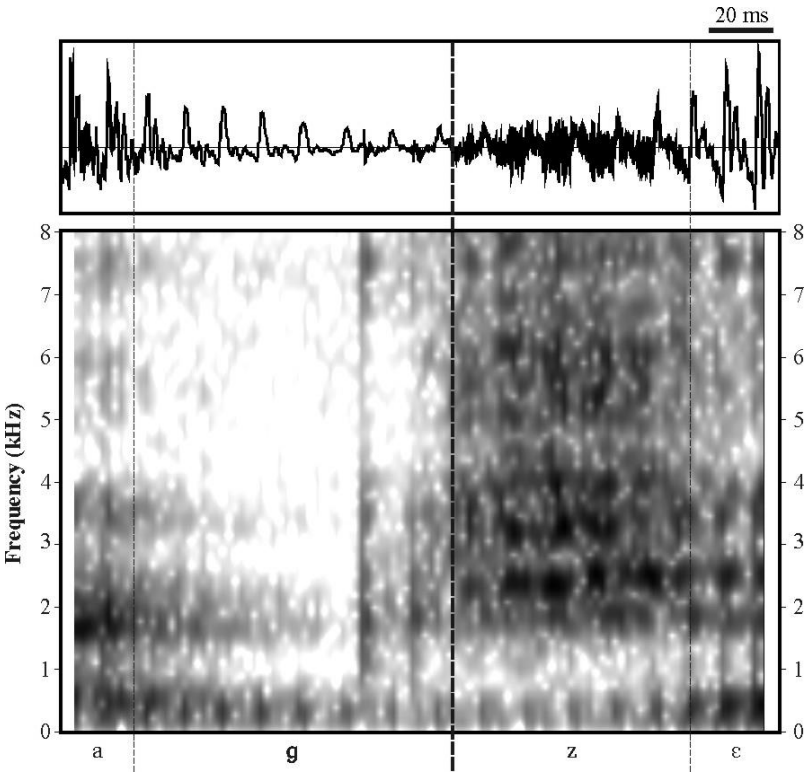


Figure 8.3. Sequence [agzɛ] with the boundary between [gz] clearly indicated by the noise.

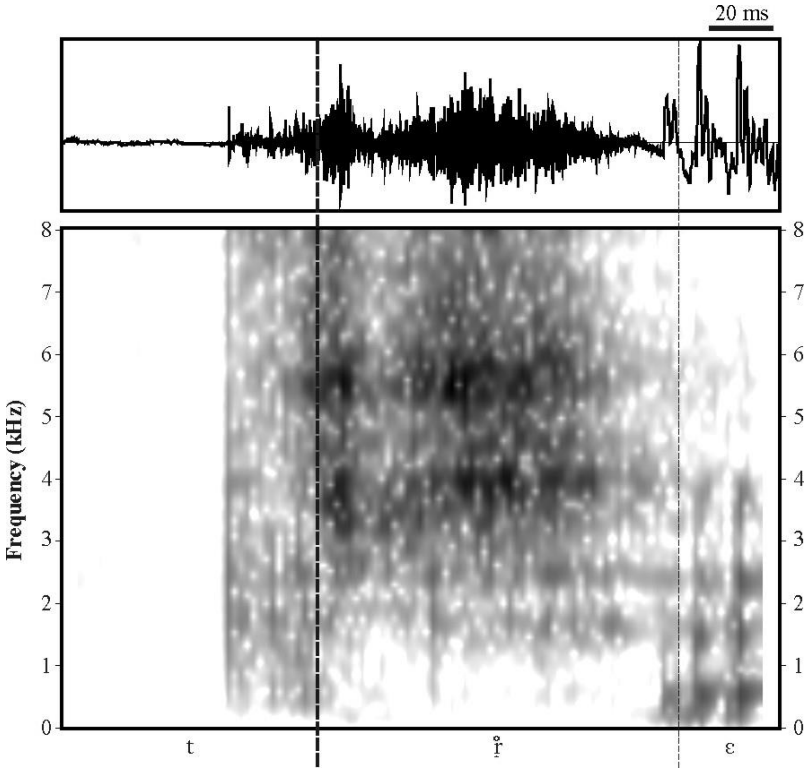


Figure 8.4. Sequence [t̥r̥ɛ] with the boundary between [t̥r̥] clearly indicated by the noise.

8.3. Additional segmentation guidelines

When segmenting fricative-plosive (or fricative-affricate) sequences, we frequently encounter the continuation of noise into the closure phase of the following plosive (*cf.* voicing continuation into the voiceless closure of an obstruent in Chapter 2). This residual noise is, in comparison with the full fricative noise of the preceding speechsound, rather weak, and it will be considered as part of the following plosive, as indicated in Figure 8.5. The boundary is thus placed at the offset of the salient friction.

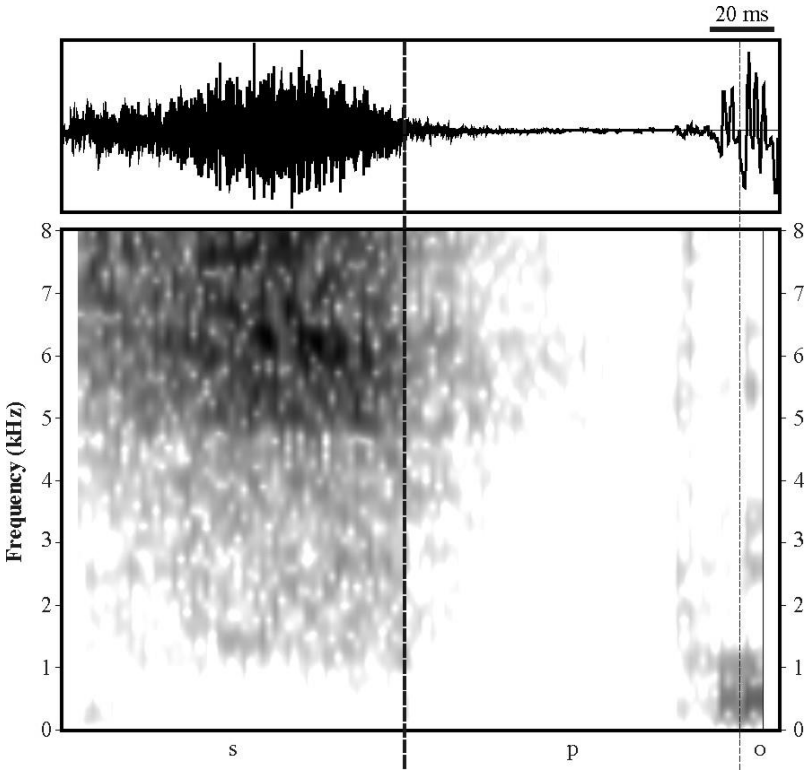


Figure 8.5. Sequence [spo] illustrating the continuation of noise in the closure phase of the plosive.

Naturally, it may happen that the decay of noise is gradual. In such cases, locating the boundary may be more difficult, and it will be necessary to identify a transition phase; the boundary will then be placed at its midpoint. The boundaries of the transition phase are defined by the beginning (or end) of full-fledged noise and the end (or beginning) of residual noise. This is illustrated in Figure 8.6 for a fricative-plosive sequence and in Figure 8.7 for a plosive-fricative sequence.

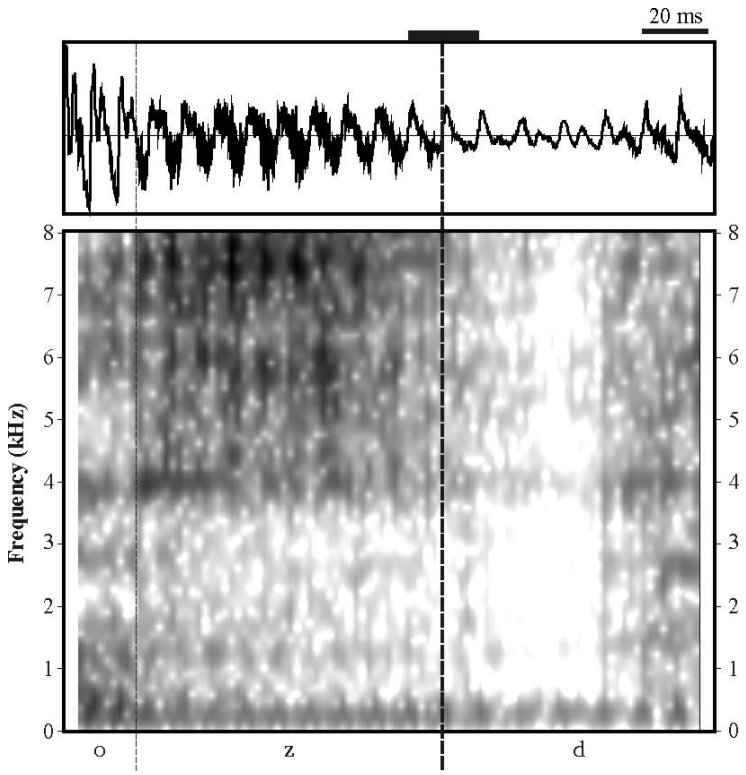


Figure 8.6. Sequence [ozd] with the boundary located at the midpoint of the transition area, which is defined as the beginning and end of noise decay.

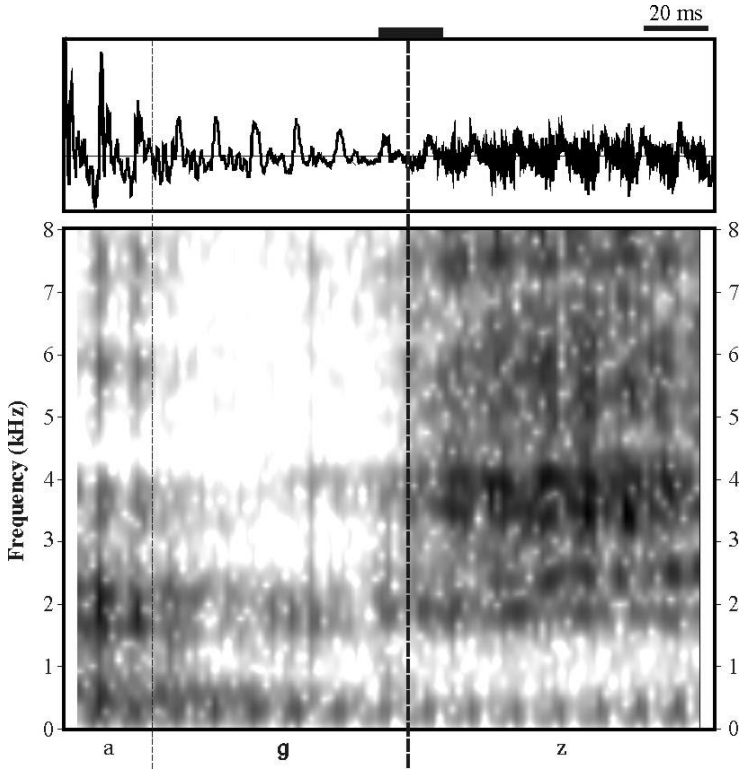


Figure 8.7. Sequence [agz] with the boundary located at the midpoint of the transition area.

It has been mentioned in Chapter 3 that (not only) the Czech /v/ behaves partly as a sonorant and partly as an obstruent. In standard Czech, /v/ does not bring about voicing assimilation, so that we may obtain a sequence of a voiceless obstruent and a fully voiced /v/. We have also mentioned that, especially in some positions, /v/ tends to have the character of a labiodental approximant rather than a fricative. An example of a sequence of an obstruent followed by /v/ is shown in Figure 8.8; the boundary is then placed at the onset of the periodic component of the sonorant /v/.

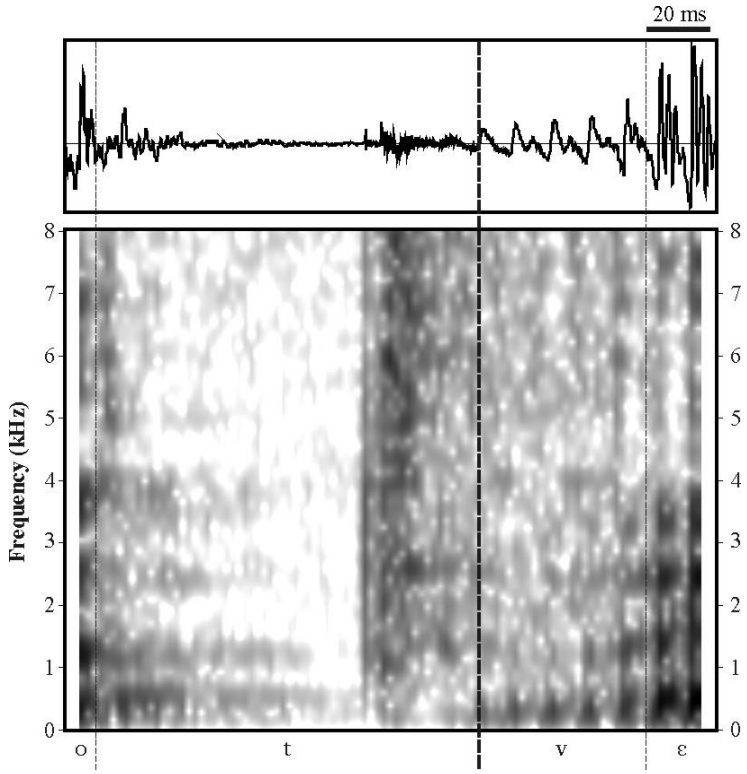


Figure 8.8. Sequence [otvɛ] with the boundary between [tv] placed at the beginning of the first clear period in the waveform.

The boundary between a fricative and a following plosive may be less obvious in the signal in items where the hold phase of the plosive also contains noise, in other words when the plosive is spirantized. In such cases, we have to look for the end of the full noise of the fricative, as indicated in Figure 8.9.

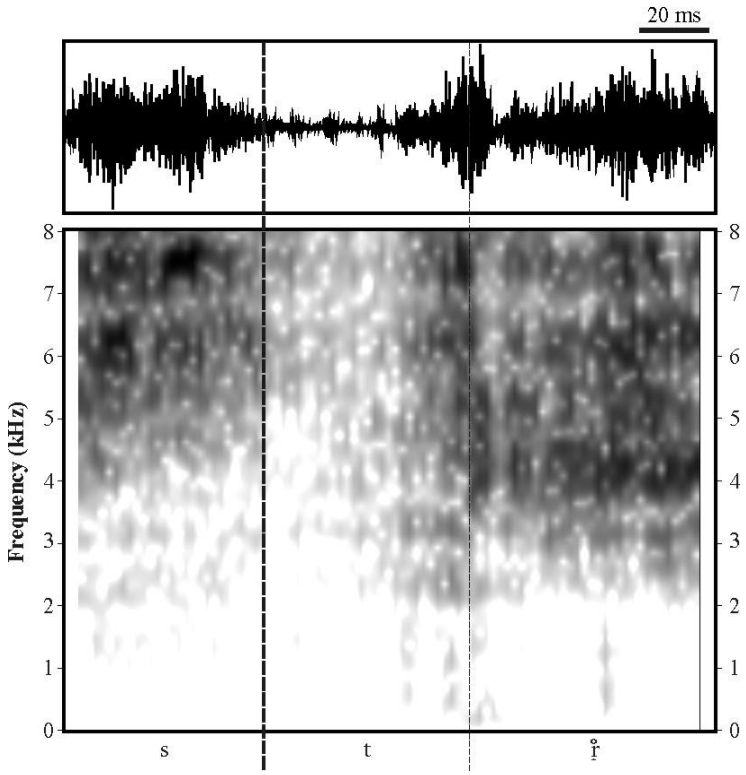


Figure 8.9. Sequence [str̥] in which [t] is spirantized; the boundary between [st] is placed at the offset of full fricative noise.

8.4. Summary

In consonant clusters containing a fricative, segmentation appears to be easiest when we use the onset and offset of full fricative noise as the primary criterion. In items with gradual onsets or offsets, we apply the rule placing the boundary at the midpoint of the transition phase. The continuation of noise in the closure phase of plosive is considered here to be equivalent to voicing continuation, that means as part of the following plosive or affricate.

9. Obstruent-liquid sequences

In many of these sequences (e.g., [br kl vr sl]), we will be able to exploit phonetic features that have been mentioned in the previous chapters for differentiating the individual types of consonants from neighbouring vowels. This includes mainly the presence of full formant structure. Figures 9.1 and 9.2 show examples where segmentation is relatively straightforward, the former with [l] and the latter with the trill [r].

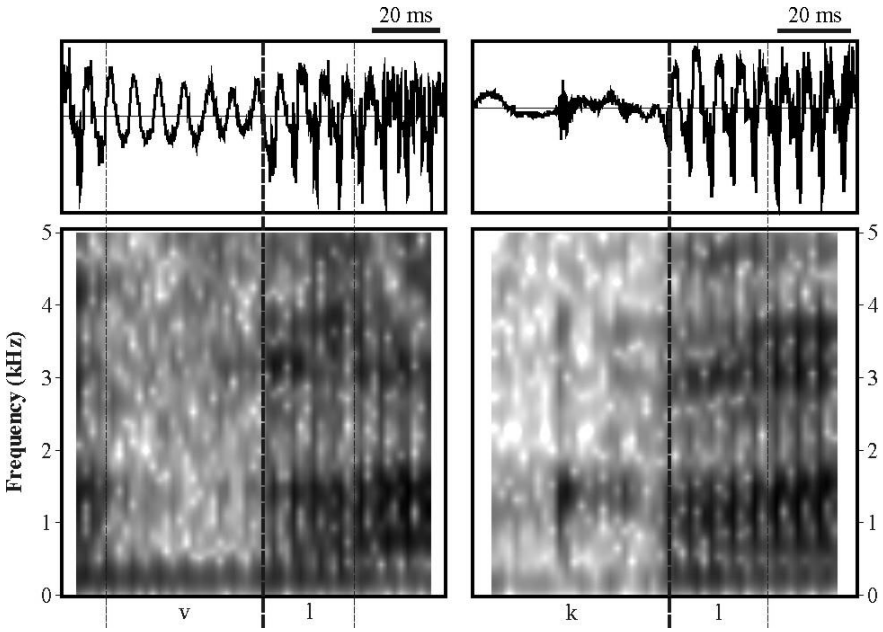


Figure 9.1. Sequences [vl] (left) and [kl] (right) with straightforward boundary placement.

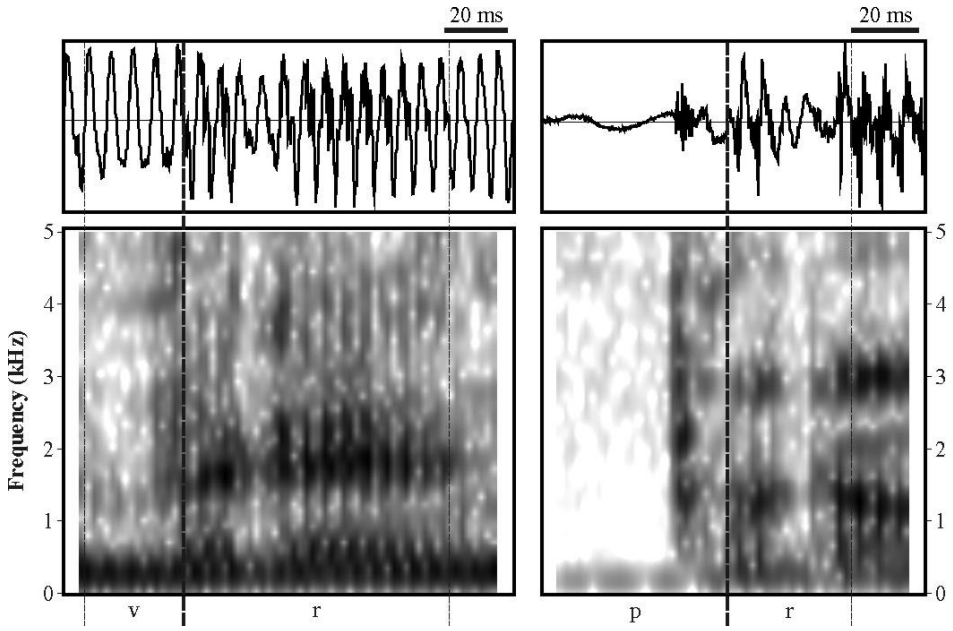


Figure 9.2. Sequences [vr] (left) and [pr] (right) with straightforward boundary placement.

In the following two sections, we are going to examine several further cases which characterize the behaviour of liquids in consonant clusters and which are relevant for segmentation. Section 9.1 will deal with clusters containing [l], section 9.2 with those containing [r].

9.1. Clusters with [l]

The specific situation regarding [l] occurs especially in sequences of a voiceless alveolar obstruent and [l]. In these (but not only these) contexts, [l] tends to become partially or completely devoiced.

Segmentation can be quite problematic in sequences of [s̥l̥] where we are essentially dealing with a sequence of two alveolar fricatives. The difference between them consists only in stridency so that intensity should be higher in [s] than in [l̥], as indicated in Figure 9.3. In some items, the visible contrast is very low and it is necessary to listen to the sequence (see Figure 9.4).

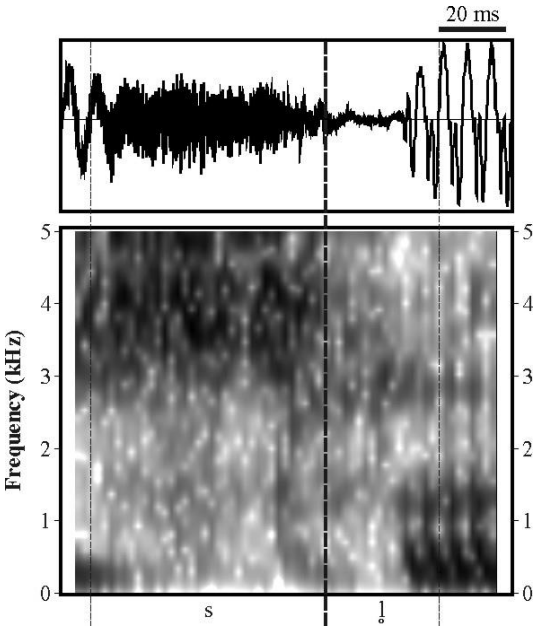


Figure 9.3. Sequence [s̥l̥] with the boundary between the two noise segments clearly signalled by the intensity of high-frequency noise.

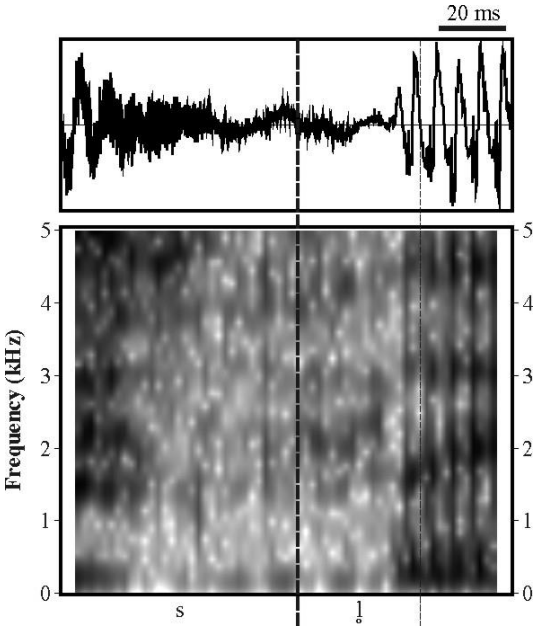


Figure 9.4. Sequence [s̥l̥] where the boundary between the two noise segments has to be determined with the aid of auditory cues.

It is surprising that devoicing of [l] occurs even with a non-alveolar fricative antecedent. Since the acoustic properties of post-alveolar and velar fricatives are different from those of [l̥], segmentation tends to be straightforward in such items.

Coming back to [sl̥] sequences, however, slight changes in the overlap of articulatory gestures can result in the presence of an epenthetic sound. This occurs in sequences of /sl/ (and also /sn/), and the epenthetic sound resembles that of [t], thus [stl̥]. As this can be regarded as a parasitic, probably unplanned sound, the epenthetic element is not segmented as a separate speechsound, and will be considered as part of the fricative [s]. Figure 9.5 shows an example of this, with the epenthetic element indicated by the arrow.

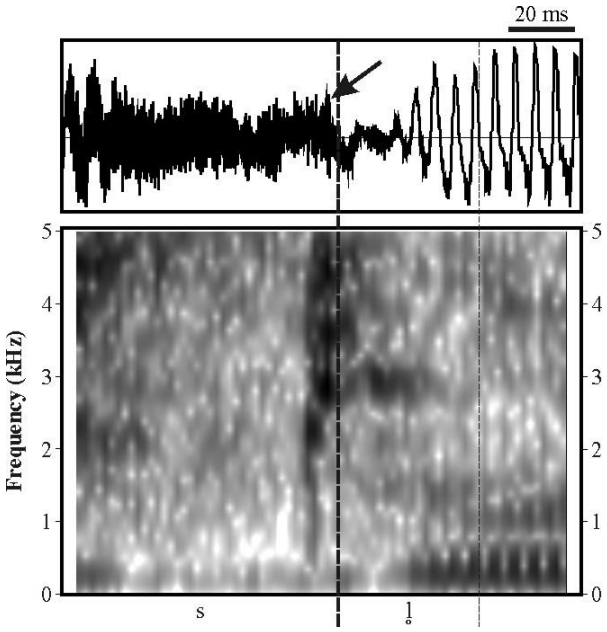


Figure 9.5. Sequence [stl̥] containing an epenthetic [t]-like element (indicated by the arrow) which is regarded as part of [s].

In sequences of a voiceless alveolar plosive with [l], we frequently encounter partial devoicing of [l] and, more importantly, lateral release of the plosive. This means that the alveolar closure is preserved, and the speaker only lowers the sides of the tongue to allow the escape of air.

To separate the plosion of [t] from the following (partially) voiceless lateral noise may be quite challenging; one could say that from the phonetic perspective the sequence [tl̥] is an affricate. It is very difficult in phonetic affricates (compare the English sequence [tɹ̥] in the next section) to separate the "plosive" part from the "fricative" part. Listening does not help in these sequences, because we tend to hear both components (both plosion and friction) for quite a long interval. Since one of the main aims of our segmentation guidelines is easy applicability and consistency, we will place the boundary between the laterally released [t] and the partially devoiced [l̥] at the onset of formant structure of the voiced part of /l/, as shown in Figure 9.6. The release stage of /t/ can therefore be quite long and caution must be exercised when we are interested, for example, in segment durations.

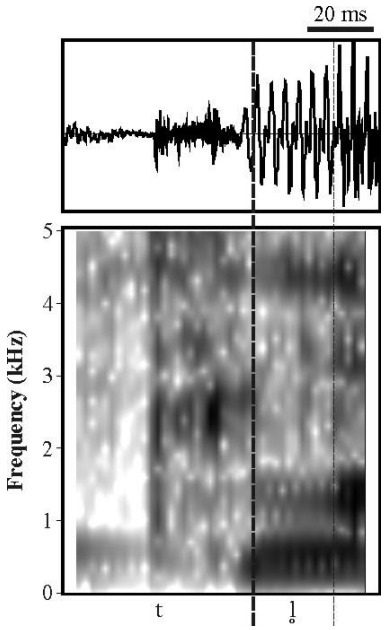


Figure 9.6. Sequence [tl̥] illustrating the segmentation of the lateral noise as part of [t].

However, the lateral noise may be clearly demarcated in the signal, and its end may not overlap with the onset of full formant structure. In such items, we do place the boundary at the end of the full noise, as shown in Figure 9.7.

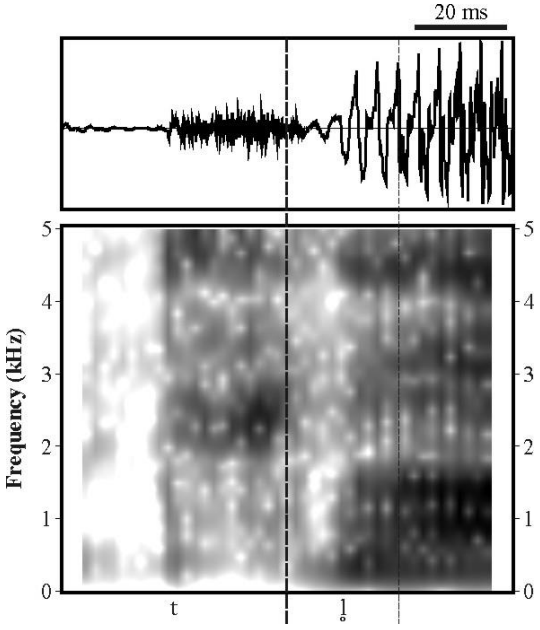


Figure 9.7. Sequence [tl̥] with no overlap between the end of lateral noise and the onset of formant structure.

9.2. Clusters with [r]

As we have mentioned in Chapter 5, the trilled [r] consists of the cycle(s) and vocalic elements. These will be apparent in consonantal clusters (*cf.* Figures 5.1 and 9.2). Thanks to these vocalic elements, boundary location between an obstruent and the following [r] does not tend to cause difficulties. Problems can arise when /r/ is not realized as a sonorant trill but as, for example, a devoiced or a fricative sound. Several interesting cases are shown in Figures 9.8 and 9.9. With such

realizations of /r/, segmentation proceeds in accordance with rules defined for fricatives in Chapter 3.

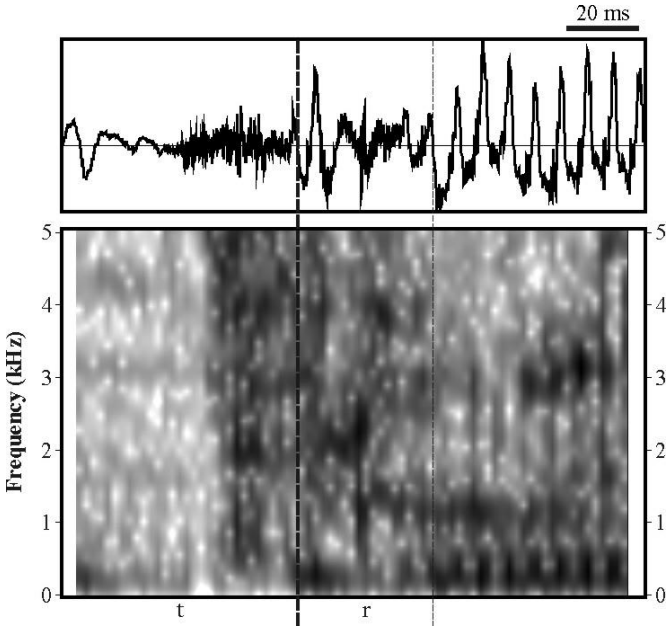


Figure 9.8. Sequence [tr] in which the waveform aids segmentation more than the spectrogram.

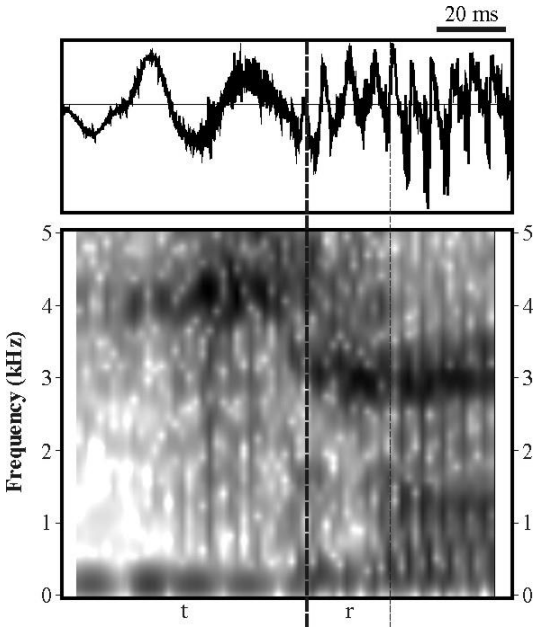


Figure 9.9. Sequence [tr] in which the character of /r/ resembles that of a voiced fricative. The contrast in the spectral composition of the noise aids segmentation here.

9.3. Summary

In many instances of obstruent-liquid sequences, we can exploit rules mentioned in chapters on intervocalic consonants. Frequently, however, /l/ becomes devoiced after voiceless obstruents, especially after [s] and [t]. In [s̥l] sequences, we have to search for the contrast between two types of noise in high frequencies. Epenthetic [t]-like elements are regarded as part of [s]. In [t̥l̥] sequences, the boundary is placed, for the sake of simplicity, at the beginning of full formant

structure of the voiced part of [l] (unless the offset of lateral noise and onset of formant structure no not coincide).

10. Sequences of speechsounds with the same manner of articulation

This chapter will examine clusters which comprise two consonants of the same manner of articulation. Such clusters can be realized in several ways, and many of the realizations will present no particular challenge from the segmentation viewpoint. However, some of them do require specific decisions to be made and appropriate rules formulated. The following sections will present guidelines for segmenting clusters with two plosives, two nasals, or two fricatives.

10.1. Clusters of two consecutive stops

The first possible realization is simply the succession of two plosives, that is, two closure-release sequences. Figures 10.1 and 10.2 show examples with a sequence of voiced and voiceless plosives, respectively. Segmentation is straightforward in these canonical realizations.

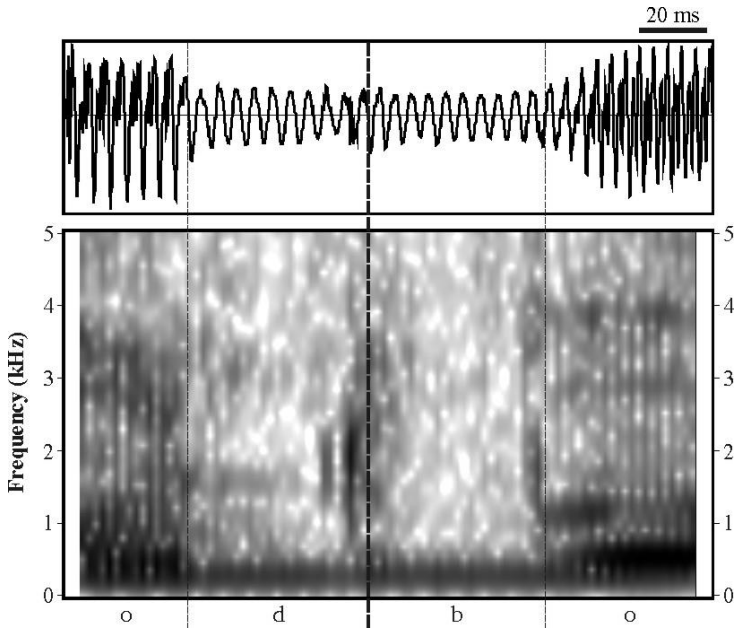


Figure 10.1. Sequence [odbo] with a canonically released [d].

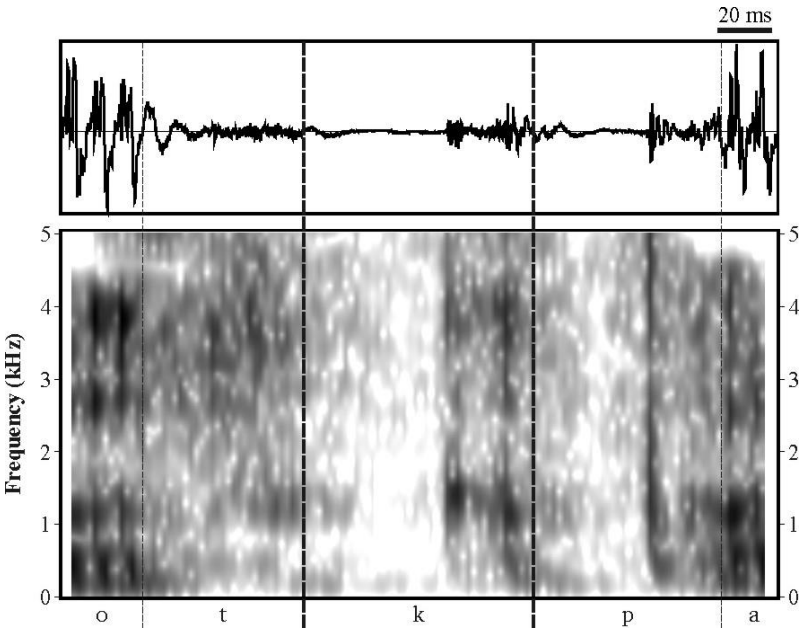


Figure 10.2. Sequence [otkpa] with incomplete closure in [t] and fully released [k] and [p].

It is possible for a short epenthetic *schwa*-like sound to appear between the two voiced plosives (e.g., *bad dog* realized as ['bædᵊ 'dɒg]). We can regard this schwa as some sort of fortification of the first plosive (*cf.* Matoušek *et al.*, 2009), and it will therefore be segmented with the first plosive. However, this epenthetic *schwa* may sometimes be so long that an additional syllable is perceived (this seems to occur often in Czech TV and radio broadcasters; for more detail see Machač & Skarnitzl, 2009). In such a case, we may want to mark it as an independent speechsound. The two options are indicated at the top and bottom of Figure 10.3, respectively.

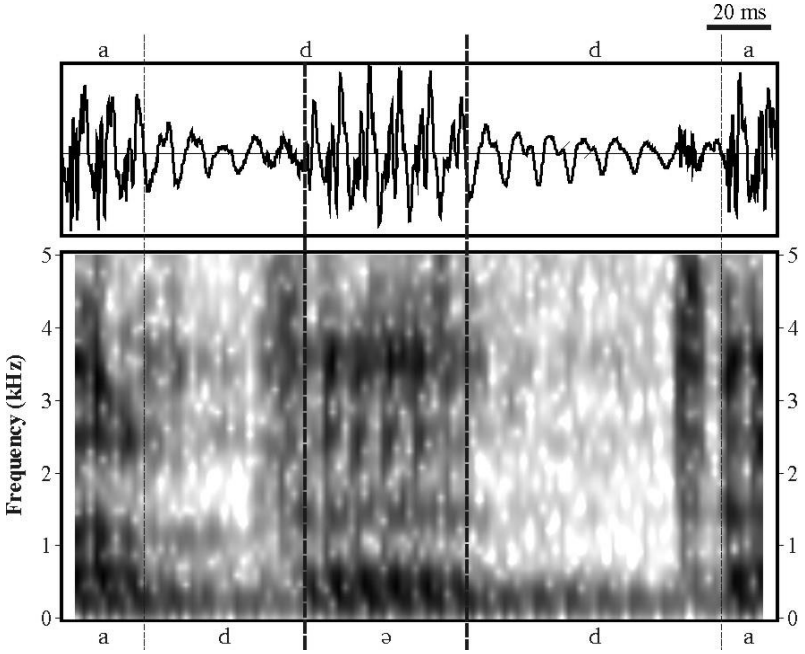


Figure 10.3. Sequence [adəda] with the epenthetic element segmented either as part of the [d] (indicated at the top) or as an independent speechsound (bottom).

Finally, the first plosive in the sequence may be unreleased. In other words, its release phase (the plosion) is missing, resulting in an aggregation of the two hold phases into a longer one. Unlike cases when the first speechsound is completely elided, the occlusion is more or less comparable in its duration to that of two sounds. Moreover, we can hear (if not see) the transition to the place of articulation of the first plosive. Obviously, there is no telling where the end of the first unreleased plosive is, so it is necessary to stipulate a rule which will not be based on any visible or audible clues. For this situation, it seems to be most sensible to place the boundary near the midpoint of the double occlusion, as indicated in Figure 10.4.

Similar situations can arise in sequences of two consecutive nasal sounds because like plosives, nasals are also stops. Figure 10.5 shows a canonically produced sequence [mɲ] with a visible release of [m] and also a slightly different spectral composition of the two nasal sounds. In Figure 10.6, we can see an epenthetic *schwa* between [n] and [m]. The epenthetic element is shorter here and does not seem to invoke the percept of an additional syllable; that is why we decided to segment it as part of the first nasal. In items of two nasals with the first one unreleased, the boundary will be placed simply near the midpoint of the long nasal sound.

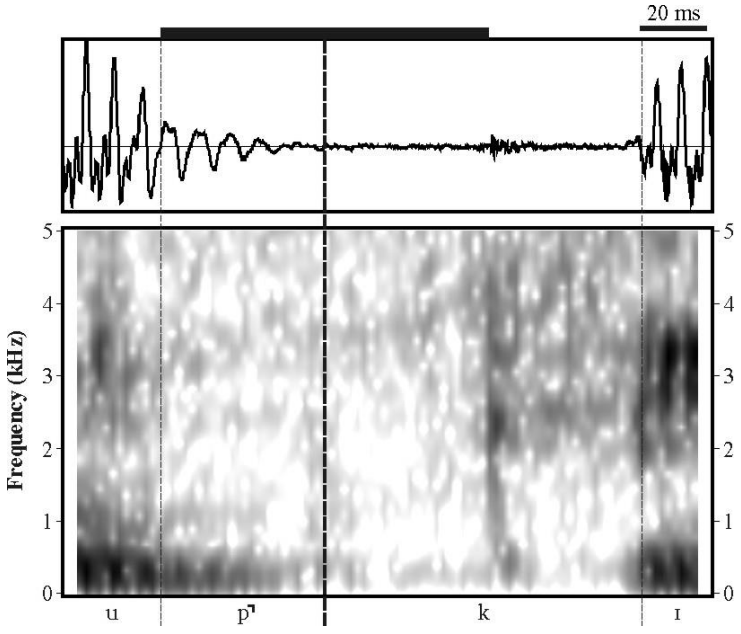


Figure 10.4. Sequence [upˀkɪ] with the boundary placed near the midpoint of the double closure. The duration of the closure is indicated by the horizontal bar at the top.

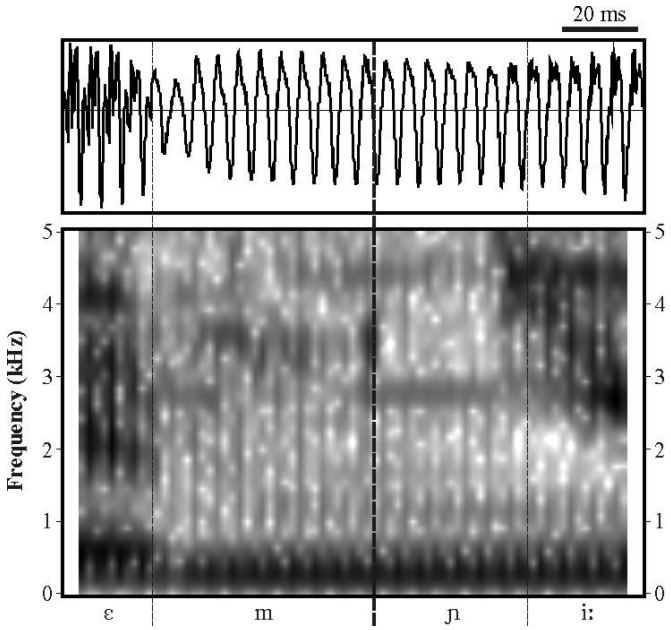


Figure 10.5. Sequence [ɛmɲiː] with the two nasal sounds quite clearly separated in the spectrogram.

10.2. Clusters of two consecutive fricatives

In a sequence of two fricatives, it is their place of articulation which is important from the segmentation perspective, more specifically whether it is the same or not. If the fricatives are of different place of articulation, spectral properties (the broadband noise formants and their relative intensity) should give indication as to the placement of the boundary. Figure 10.7 gives one such example with two consecutive voiceless fricatives, Figure 10.8 another with two neighbouring voiced fricatives. The spectrograms in this section display the frequency range between 0 and 8 kHz (*cf.* Chapter 3 on fricatives).

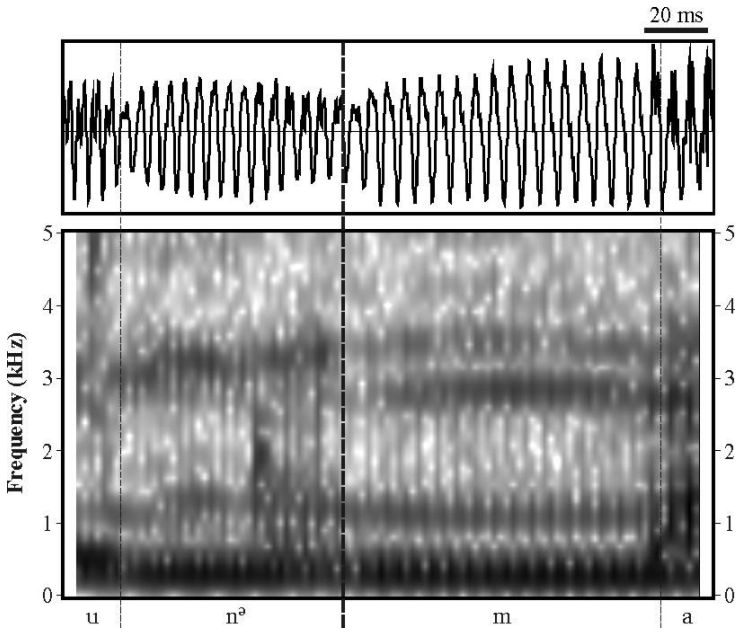


Figure 10.6. Sequence [unᵊma] with the epenthetic *schwa* segmented as part of the first nasal. (See Chapter 4 for guidelines regarding the boundary between a vowel and a nasal.)

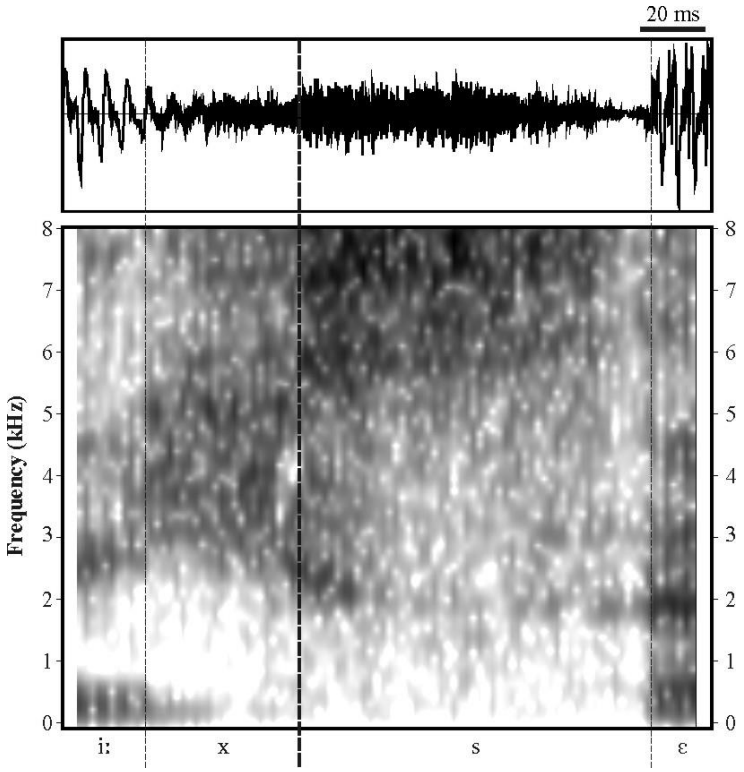


Figure 10.7. Sequence [iːxsɛ] in which the boundary between [x] and [s] is indicated by the difference in intensity, especially in high frequencies. Note that /x/ is fronted here, [x̟], and the noise formant thus lies higher than in a truly velar [x].

Unfortunately, the contrast in the frequency and/or intensity of noise formants may not always be so clear as to allow unambiguous segmentation. In some items, we will therefore have to exploit listening and possibly also the rule of placing the boundary near the midpoint of a transition area. Figure 10.9 illustrates such a case for the sequence of [s] and [ʃ].

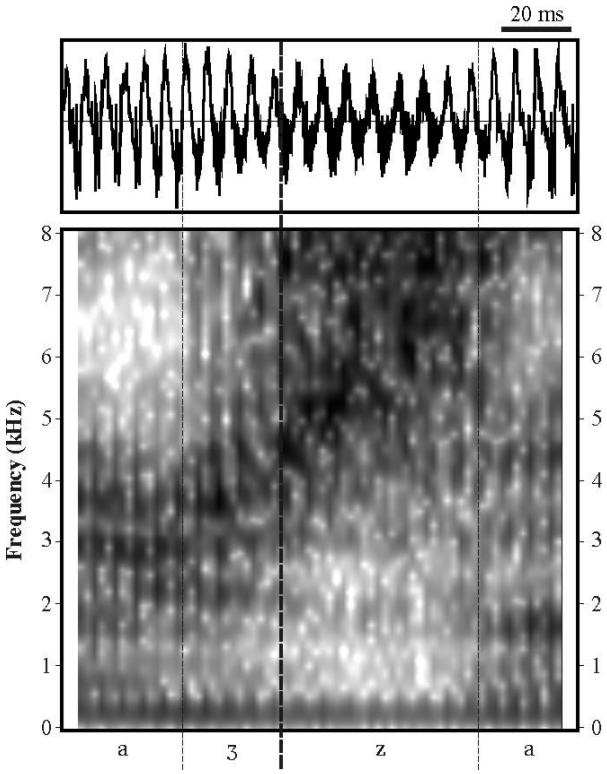


Figure 10.8. Sequence [aʒza] where the boundary is indicated by the difference in frequency and intensity of the fricatives' noise formants.

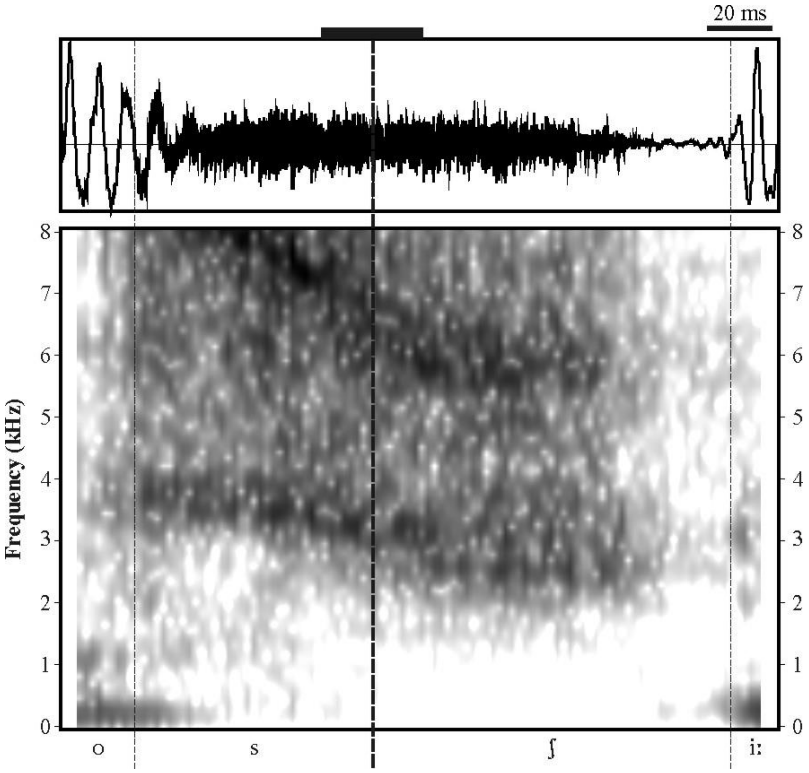


Figure 10.9. Sequence [osʃiː] in which the boundary is placed near the midpoint of the transition area (indicated by the black bar at the top).

If the two neighbouring fricatives are of the same place of articulation, segmentation is usually quite difficult. Typically, there are no visual cues separating the two speechsounds, and, as may be expected, listening does not help to separate them. It may even be difficult to decide whether we are

dealing with a “long” fricative, i.e. a geminate (a sequence of two acoustically indistinguishable fricatives), or with one single fricative (in which case the other one has been elided). Listening may help with this decision. For an example of elision in English, imagine the phrase *a couple of friends* pronounced very fast: it may be [ə ˈkʌpl̩ əf ˈfrendz] or with the elision of the [f] in the preposition, [ə ˈkʌpl̩ ə ˈfrendz].

Figure 10.10 shows a sequence taken from the Czech word *rozsáhlý* /ˈrossaːhliː/ (*vast*). The fricative sound is approximately 120 ms long, which theoretically may correspond to one or two speechsounds. Based on local speech rate and also auditory analysis, we treat it as a sequence of two sounds. In such a case, there is no other option but to place the boundary between the two [s] components into the temporal midpoint of the entire fricative sound.

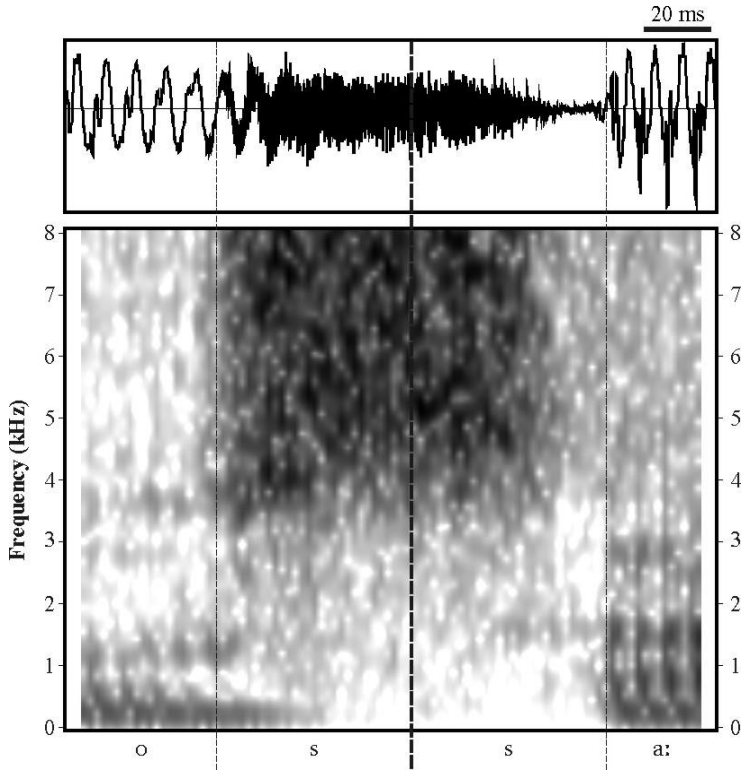


Figure 10.10. Sequence [ossaː] in which the boundary between the two [s] constituents is placed in the middle of the entire fricative sound.

In Section 10.1, we have mentioned the possibility of an epenthetic element appearing between both two plosives and two nasals. The same can happen with fricatives. Segmentation is similar in these situations as with plosives and nasals: if the presence of epenthetic *schwa* results in the impression of an additional syllable, it may be considered as an independent speechsound. Otherwise, it is regarded to constitute a reinforcement of the first fricative and thus segmented with the first fricative (see Figure 10.3 for the two options). Figure 10.11 shows the sequence of two [z] sounds from the phrase *dnes začnou* (*today begin* [they]); since the vocalic element is over 60 ms long and an extra syllable is quite clearly heard, it is segmented as a separate speechsound, *schwa*.

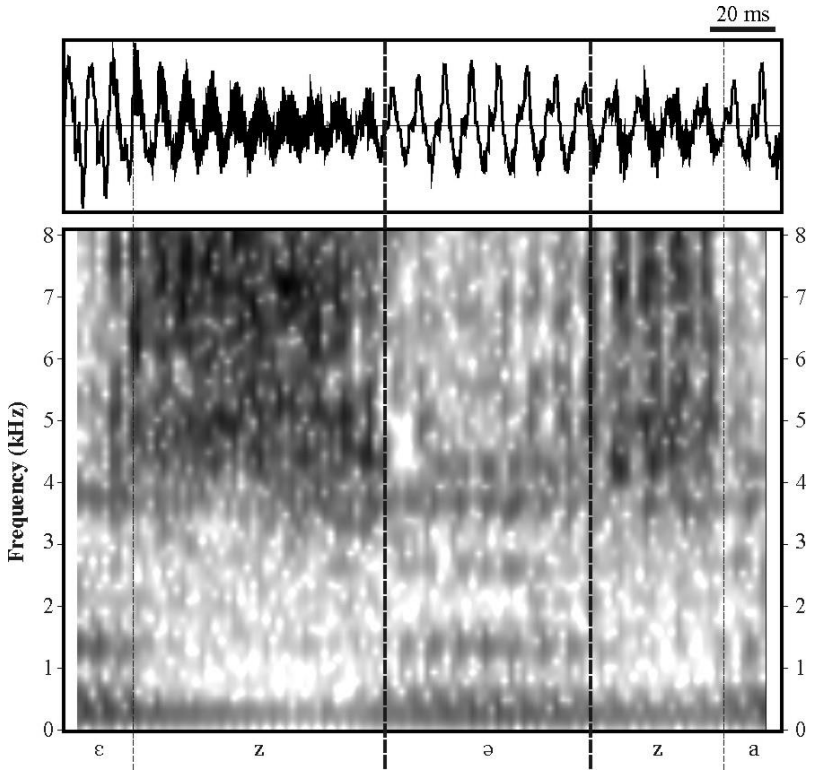


Figure 10.11. Sequence [ɛzəza] with an epenthetic *schwa* regarded as an independent speechsound.

10.3. Summary

In sequences of two sounds with the same manner of articulation, there are typically three options. We may encounter canonical realization: plosives and nasals will be released, and there will be a clear separation of two consecutive fricatives. This is very rare in fricatives with the same place of articulation, though. With canonical realizations, we exploit the segmentation rules presented in the relevant chapters.

In realizations which may be called "unreleased", the boundary is typically placed near the midpoint of the acoustically compact sound (i.e., the closure in plosives, the nasal murmur in nasals, and the fricative portion in fricatives).

Finally, an epenthetic vocalic element between the two speechsounds may be regarded as part of the first sound (reinforcing it) or as a separate speechsound, depending largely on the impression of syllabicity and possibly the purpose of segmentation.

11. The glottal stop in word-initial vowels

In this chapter, we will examine the segmentation of the phonological category called the **glottal stop [ʔ]**. The glottal stop is a speechsound which often precedes word-initial vowels in languages such as Czech or German, in which linking (liaison) is not as frequent as in English or French. Phonetically, the glottal stop may assume several forms, by far the most frequent ones being a canonical plosive and creaky voice (Redi & Shattuck-Hufnagel, 2001; Skarnitzl, 2004). The

following two sections will present segmentation guidelines for the plosive-like glottal stop and for the creaky glottal stop, respectively.

11.1. Plosive-like glottal stop

In instances of canonical glottal stops, whose articulation consists simply of two stages, the closure phase and the plosion, segmentation is relatively straightforward. Figure 11.1 shows such an example. The plosion of [ʔ] is easy to separate from the glottal periods of the subsequent vowel, especially thanks to the higher spectral intensity of the glottal pulse.

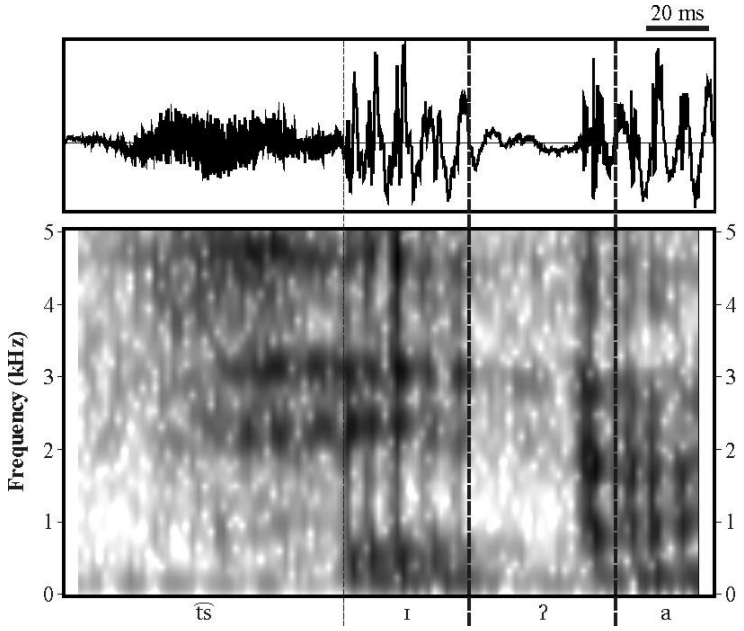


Figure 11.1. Sequence [t͡sɪʔa] with a canonical realization of the glottal stop.

However, the right boundary of the glottal stop may not always be so obvious from the signal since the glottal stop usually contains the resonances (formants) of the following vowel. In Figure 11.2, we can see one weaker period, followed by three high-intensity periods, and then the relatively regular periods pertaining to the vowel. The question is what to treat as the plosion of [ʔ]: only the first, weak period; the weak period and the subsequent three periods (which would be comparable to the solution in Figure 11.1); or a compromise between these two alternatives. As the segmentation in Figure 11.2 indicates, we have decided for the second possibility because the first four periods (the weak one and the subsequent three high-intensity periods) appear to be quite distinct, both visually and auditorily, from the subsequent periods with modal phonation.

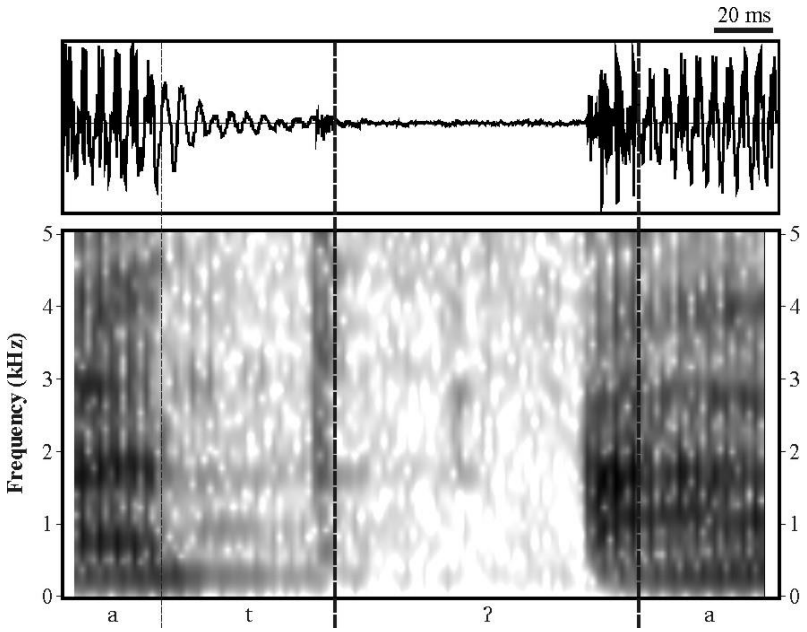


Figure 11.2. Sequence [atʔa] with a canonical realization of the glottal stop; the high-frequency intensity periods are considered, also on the basis of listening, as part of the glottal stop.

The glottal stop, while still involving a sequence of a hold phase and plosion, may also be realized in a slightly different way. The end of the preceding speechsound may already be glottalized as well, resulting in what we have called, due to its shape, the *barbell* glottal stop (Skarnitzl, 2004). A typical example is shown in Figure 11.3.

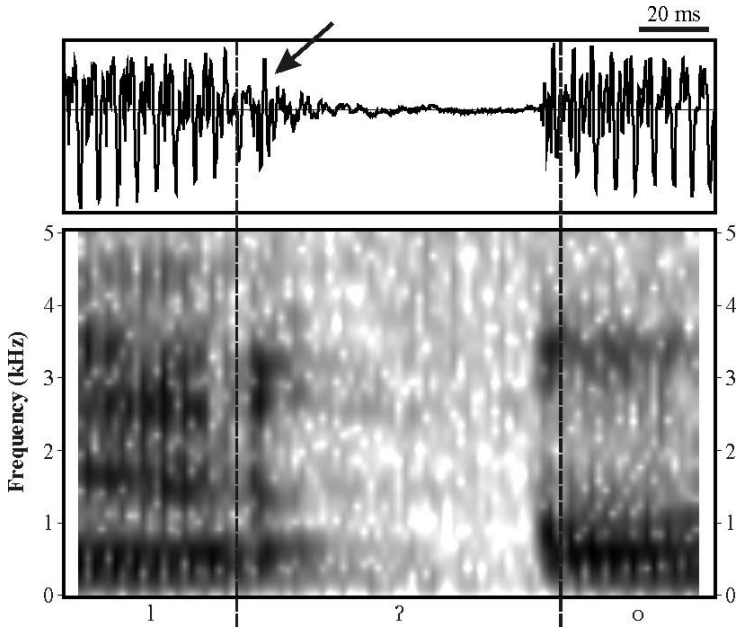


Figure 11.3. Sequence [lʔo] in which the last period before the closure (indicated by the arrow) is clearly detached from [l] and is regarded as part of glottalization in the form of a "barbell".

In other words, clearly separated glottal periods, or the onset of aperiodicity at the end of the preceding speechsound will be considered as part of the glottal stop. Figure 11.4 shows another example, which is also interesting with respect to the right boundary of [ʔ]: unlike the previous

items, the intensity of the plosion “period” is lower here than that in the subsequent vowel. This seems to occur quite frequently when the vowel in question is [u].

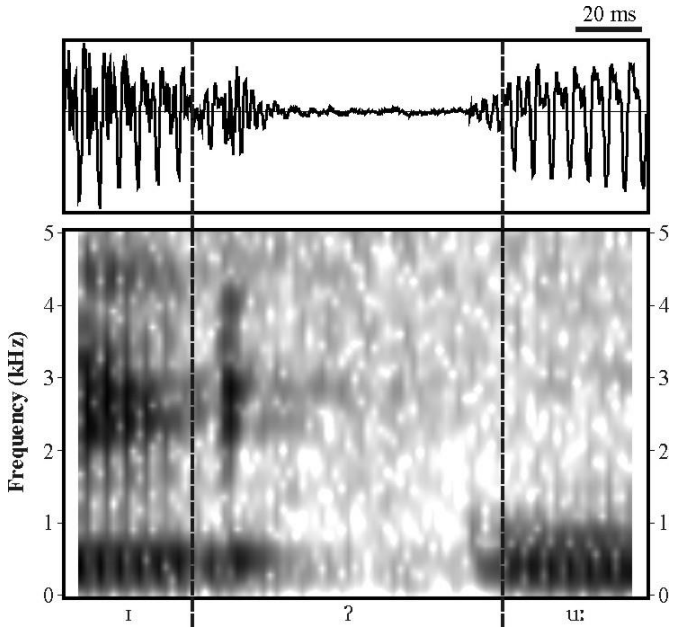


Figure 11.4. Sequence [ɪʔuː] showing a barbell glottal stop with a lower-intensity plosion of [ʔ].

11.2. Creaky glottal stop

For the purposes of this section (i.e., segmentation), a creaky realization of the glottal stop will be considered to involve some kind of irregularity in the glottal periods. We will apply the same rules as in the previous section. In other words, we are interested in specifying the point where aperiodicity of the creaky phonation starts and ends. Figure 11.5 shows an example in which the glottal stop contains only two periods, and segmentation is straightforward. In Figure 11.6, the aperiodicity is more visible in the waveform than in the spectrogram; that is probably caused by the quality of the following vowel, [u].

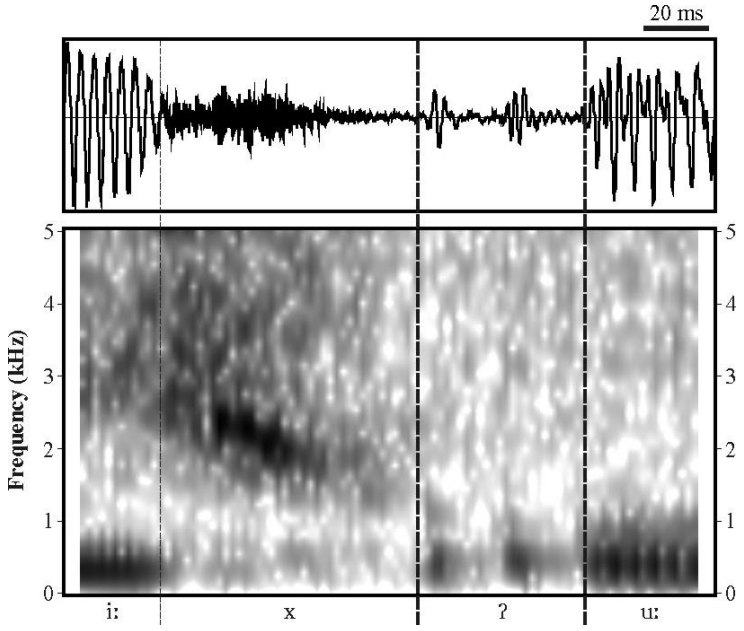


Figure 11.5. Sequence [iːxʔuː] with two aperiodic glottal pulses in [ʔ].

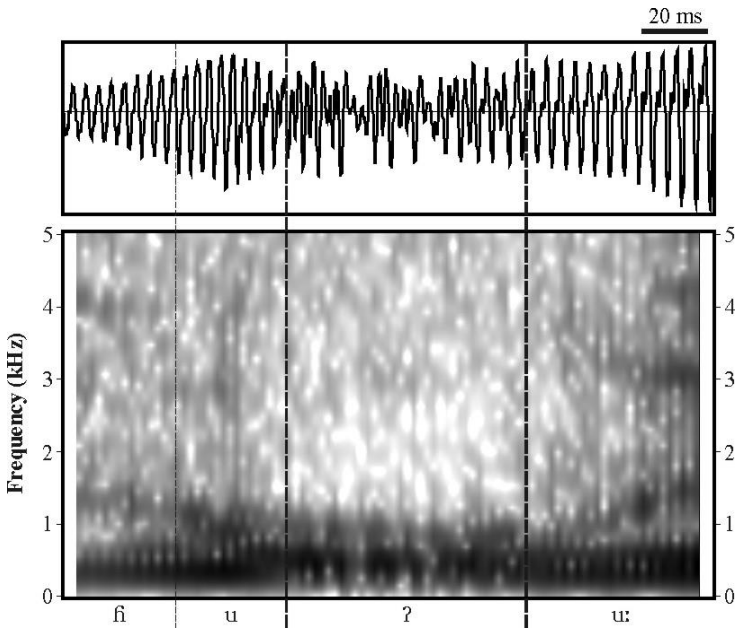


Figure 11.6. Sequence [ɦuʔuː] with aperiodicity visible especially in the waveform.

The approach adopted here, in which any aperiodicity is regarded as part of the glottal stop, has one disadvantage: the vowels neighbouring on the glottal stop (regardless of its realization) may, in the end, be very short. This is visible in Figure 11.7 and even more so in Figure 11.8 where the vowel [o] which follows the glottal stop is only approximately 20 milliseconds long.

In a study focused on temporal modelling of Czech segments, we have shown that vowels in the glottal-stop context indeed are considerably shorter than in otherwise comparable contexts (Volín & Skarnitzl, 2007). This has to be taken into account in similarly motivated research: the presence of the glottal stop seems to be an important parameter for temporal modelling, at least in Czech.

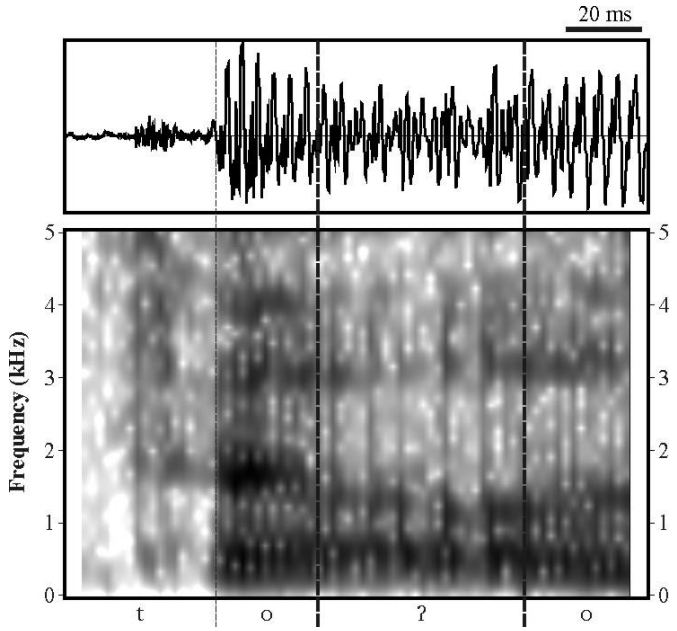


Figure 11.7. Sequence [toʔo] with a short [o] preceding the glottal stop due to the extensive aperiodic component.

It is interesting to note that the presence of aperiodicity does not appear to be necessary for a salient perception of the glottal stop (*cf.* Dilley *et al.*, 1996: 430). The sequence in Figure 11.9 has been taken from the phrase *byl italské* (*was* [*of*] *Italian*), and the presence of the glottal stop is auditorily unambiguous. We can see that Praat's autocorrelation pitch extractor can detect F0 throughout. It seems that modal phonation of very low fundamental frequency, probably along with the low amplitude of the pulses, is sufficient for the impression of a glottal stop.

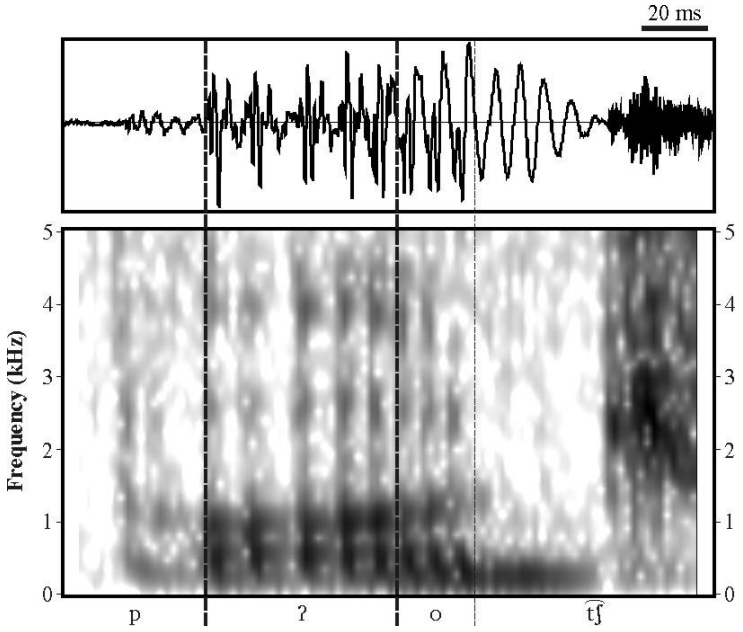


Figure 11.8. Sequence [pʔotʃ] with a very short vowel following the glottal stop.

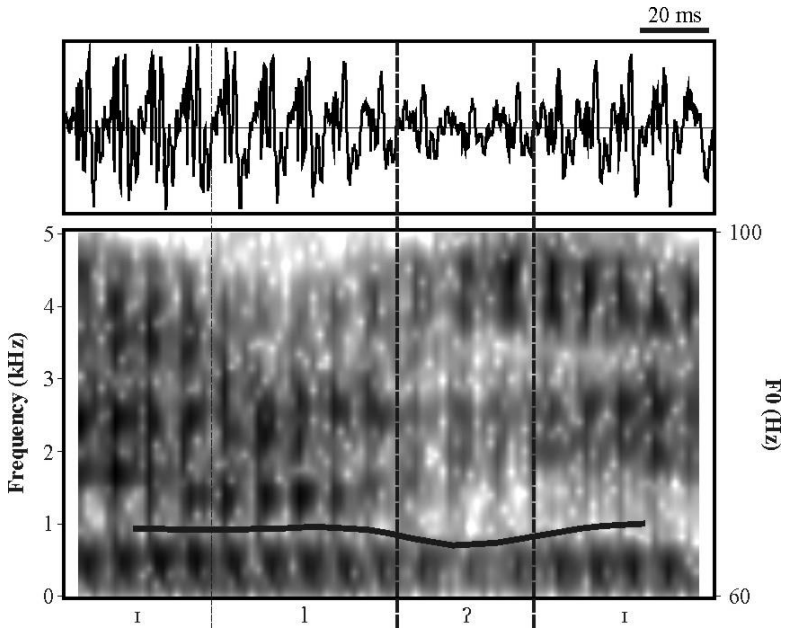


Figure 11.9. Sequence [ɪlʔɪ] with no aperiodicity, merely low F0 and lower amplitude of the pulses in the glottal stop.

11.3. Summary

The most important aspect for the segmentation of the glottal stop, whether it is a true plosive or it has the form of creaky voice, is the boundary between periodicity and aperiodicity. Since the objectives of our guidelines include easy application and comparability, aperiodic portions of the signal are segmented as part of the glottal stop. The labeller should bear in mind, however, that this may result in very short durations of neighbouring vowels.

12. Utterance beginnings and ends

The final segmentation chapter of this handbook deals with speechsound realizations at the beginnings and ends of an utterance, in other words, in the neighbourhood of a silent pause. Speechsounds may be realized differently (that is, not canonically) in pausal contexts, especially due to the specific coordination between glottal and articulatory activities. Although the underlying principles are quite similar, we will examine initial (post-pausal) and final (pre-pausal) contexts separately.

12.1. Initial speechsounds

First of all, let us look at utterances beginning with voiceless plosives or affricates. Obviously, it is impossible to identify the beginning of a voiceless plosive. We therefore need to stipulate a time interval which we will regard as the closure phase of voiceless plosives. In post-pausal contexts, the exact value is not essential; that is why we will only recommend that voiceless plosives be segmented with **closure phases of 40 to 70 milliseconds**. That corresponds to the duration of voiceless occlusions in other contexts. In the case of particularly sensitive analyses, such segments should be naturally excluded.

As we have already mentioned, the activity of the vocal folds is not completely synchronized with that of the articulating organs. Glottal activity can both precede and lag behind oral gestures. We will first examine the second case, phonation starting later than articulation. This is relevant for voiced speechsounds, both obstruents and sonorants, which may become **partially or completely devoiced** in the initial context. Czech phonologically voiced plosives should, even in initial positions, be fully voiced (unlike in English or German), with F0 present throughout the closure phase. If an initial plosive is devoiced, we will again mark 40-70 ms as the closure phase. Figure 12.1 shows the application of this rule at the beginning of the Czech word *babička* (*grandmother*) in which the /b/ is almost completely devoiced, [b̥]. The hold phase is 40 ms long here.

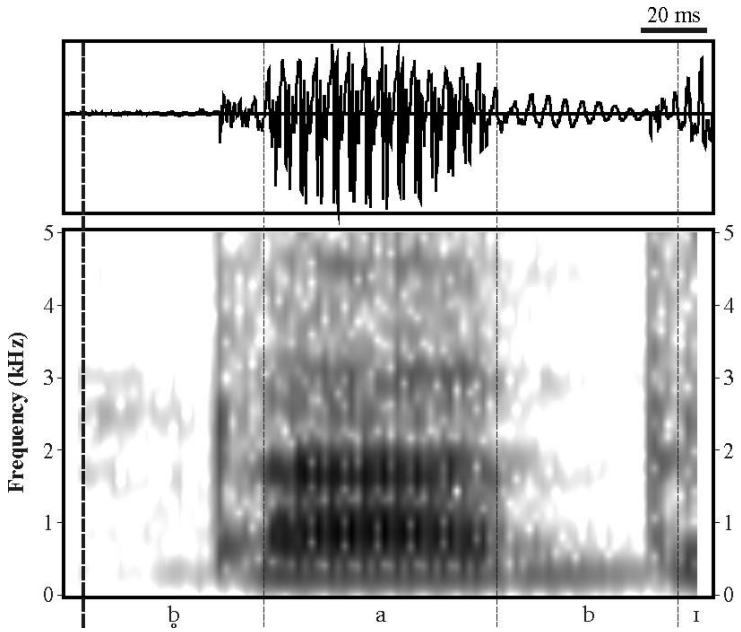


Figure 12.1. Sequence [b̥abɪ] with devoiced [b̥].

Initial devoicing can also apply to sonorants, sounds which are canonically voiced throughout. Figure 12.2 shows partially devoiced /j/ in the initial position; the voiceless palatal noise is both visible and audible. Figure 12.3 shows devoicing in an initial nasal sound, /n/. In both cases, we can see, in the frequency domain, continuity of the noise with the formants in the voiced portion of the given speechsound.

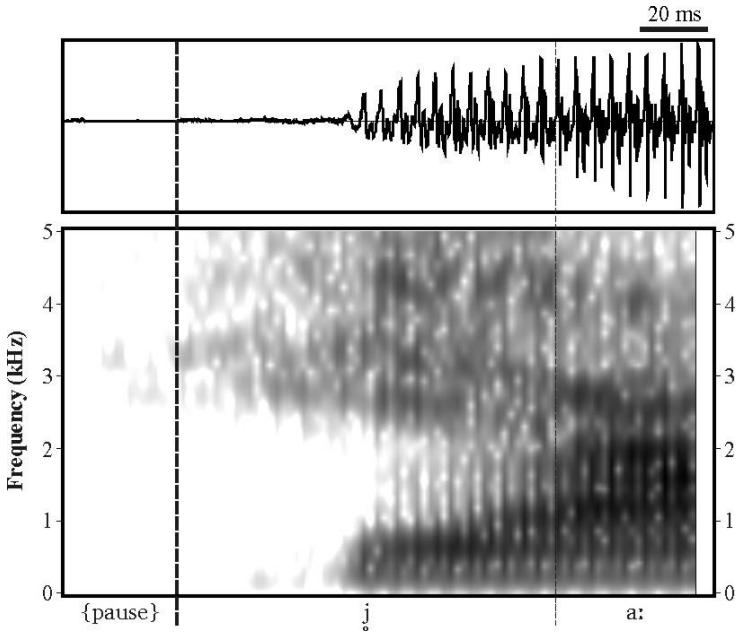


Figure 12.2. Sequence [j̥a:] with clearly visible palatal noise at the beginning of /j/.

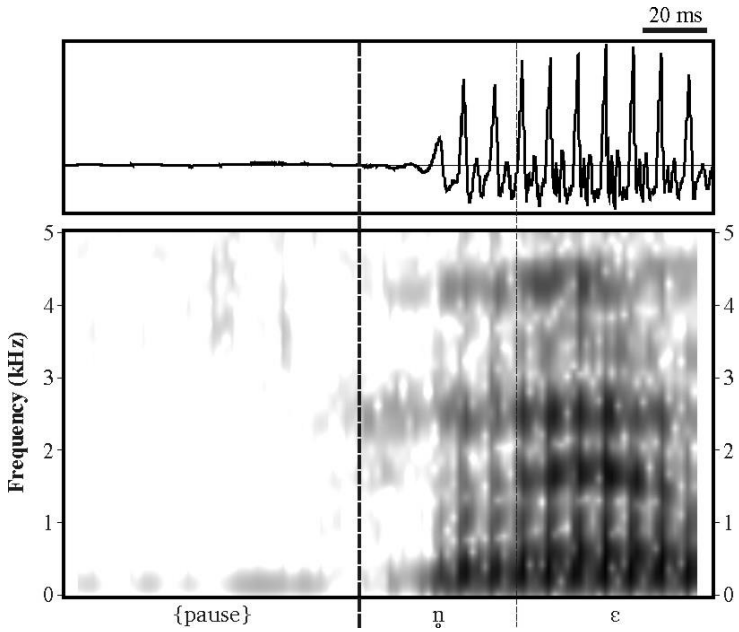


Figure 12.3. Sequence [n̥ɛ] with partially devoiced [n̥].

Initial devoicing results from the situation when glottal activity lags behind articulatory activity. In the opposite situation, the vocal folds start vibrating before the articulating organs reach their target. This has been called **preglottalization**. What is important for pregottalization is the character of vocal fold vibration, more specifically, whether it is aperiodic or periodic. We can also talk about hard glottal onset and soft glottal onset, respectively (*cf.* Machač & Skarnitzl, 2009).

Hard glottal onsets involve an abrupt, high-intensity beginning of phonation. Essentially, this is what we call the glottal stop at the beginning of initial vowels. Such glottal onsets do not typically appear before consonants but have been documented as a sort of manneristic expression of some Czech radio broadcasters (Machač & Skarnitzl, 2009). From the segmentation viewpoint, a hard glottal onset is not difficult to detect. The question remains whether preglottalization should be segmented as a specific speechsound; this will depend on the research question. Hard glottal onsets can appear before both obstruents (typically voiced, but not exclusively) and sonorants, as shown in Figures 12.4 and 12.5.

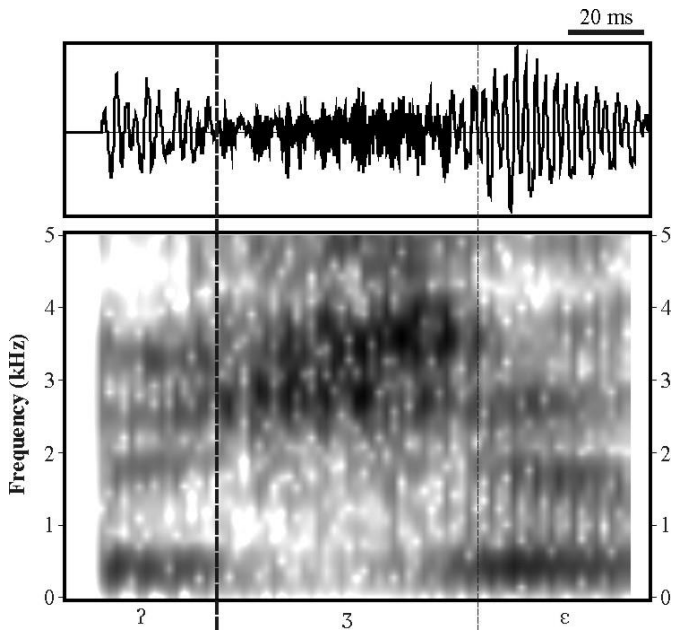


Figure 12.4. Sequence [ʔʒɛ] with a hard glottal onset before a voiced obstruent.

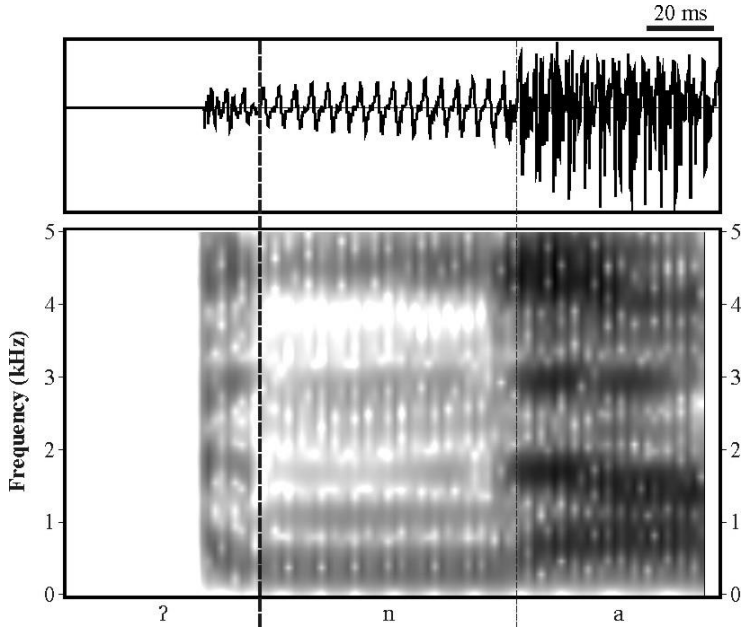


Figure 12.5. Sequence [ʔna] with a hard glottal onset before a sonorant.

Figure 12.6 shows an example of preglottalization with a **soft glottal onset**, i.e. a gradual increase of the amplitude of glottal cycles, accompanied by a *schwa*-like vocalic element. Again, this phenomenon does not present much challenge for segmentation.

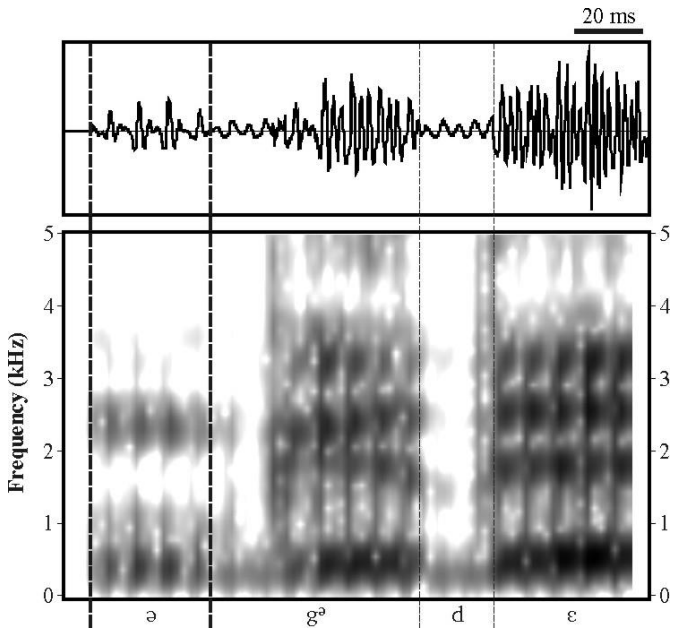


Figure 12.6. Sequence [ə'gᵊdɛ] with a soft glottal onset.

12.2. Final speechsounds

In positions before the pause, there are several ways in which phonation can end (apart from the situation when articulatory and phonatory activities are coordinated). These types of cessation of phonatory activity before the pause are relevant to vowels and sonorants. First of all, modal voicing may continue after articulatory activity has ceased. In such situations, we try follow the formant structure criterion, although the decay of formant structure may be gradual. In Figure 12.7, the final boundary is placed near the midpoint of the decay. The voicing continuation, which continues for more than 50 milliseconds, is very salient here, even auditorily.

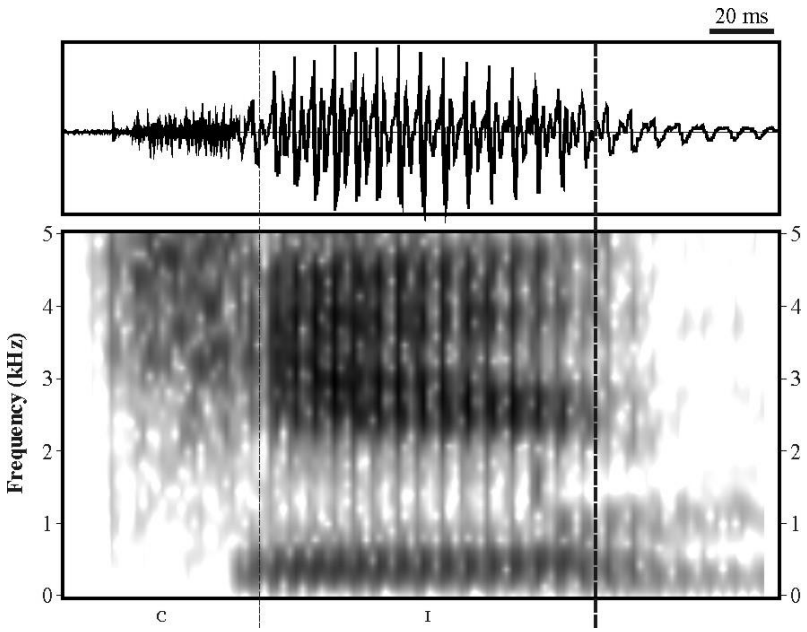


Figure 12.7. Sequence [cɪ] with long voicing continuation after the articulation of the vowel has ceased.

As a second possibility, the reduction in volume velocity of air expelled from the lungs and the relaxation of laryngeal musculature at utterance ends may lead to creaky or breathy phonation in the final utterance syllable(s). Since the amount of energy in the signal tends to be low, segmentation may be slightly more difficult, especially in breathy syllables.

Figure 12.8 shows an item with **creaky phonation** in which the end of the utterance is easy to identify. Following the formant structure criterion, the boundary will be placed at the end of the last salient formant column. In Figure 12.9, the utterance-final vowel is pronounced with quite strong **breathiness**. Listening had to be used in this case to aid visual cues because formant structure is to a large extent masked by the noise. The last 15 milliseconds, in which the noise is much stronger than formants, still have a clear [ɪ] quality, which is why they are segmented as part of the vowel.

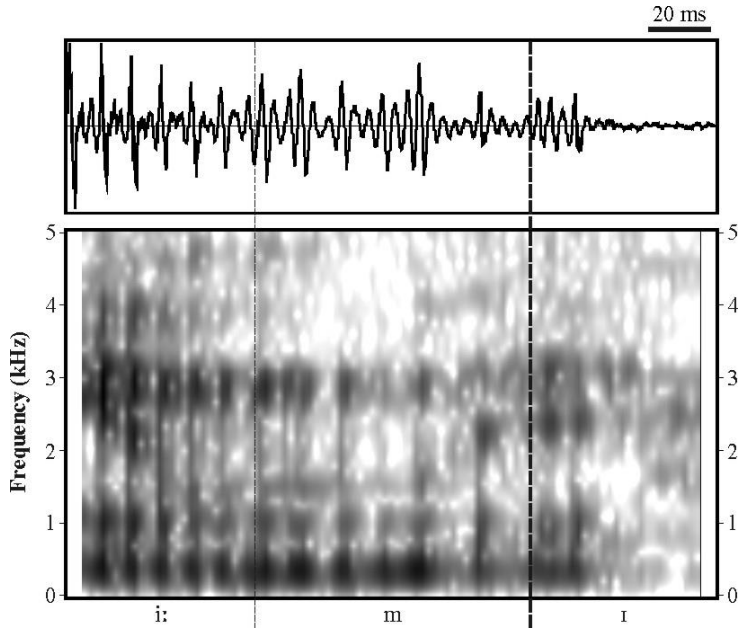


Figure 12.8. Sequence [iːm̰ɪ] with creaky phonation in the last syllable.

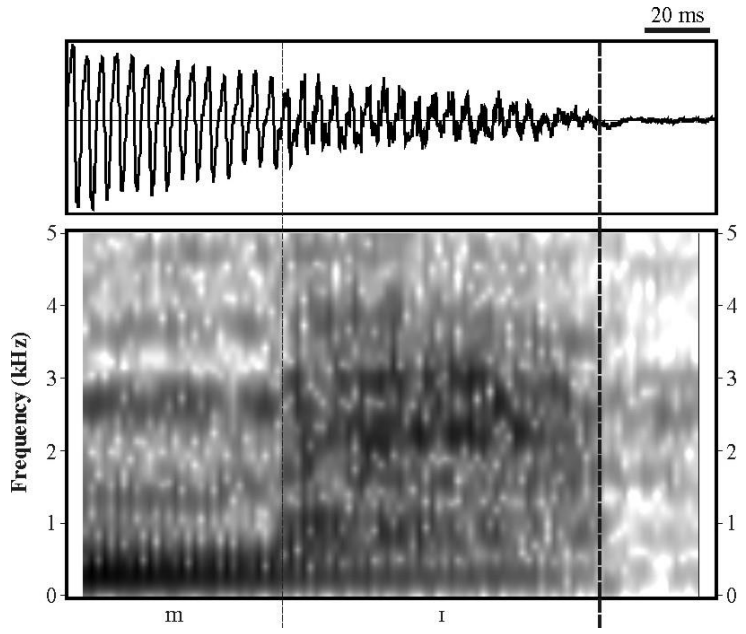


Figure 12.9. Sequence [mɪ̤] with breathiness in the final vowel.

It is interesting to note, although this will not bring anything new from the segmentation viewpoint, that creaky phonation and breathiness can appear within one segment. This is illustrated in Figure 12.10 in which the last three periods of [a], the whole [l], and about 70 percent of [ɛ] are creaky, but the last part of [ɛ] is breathy, both visually and auditorily.

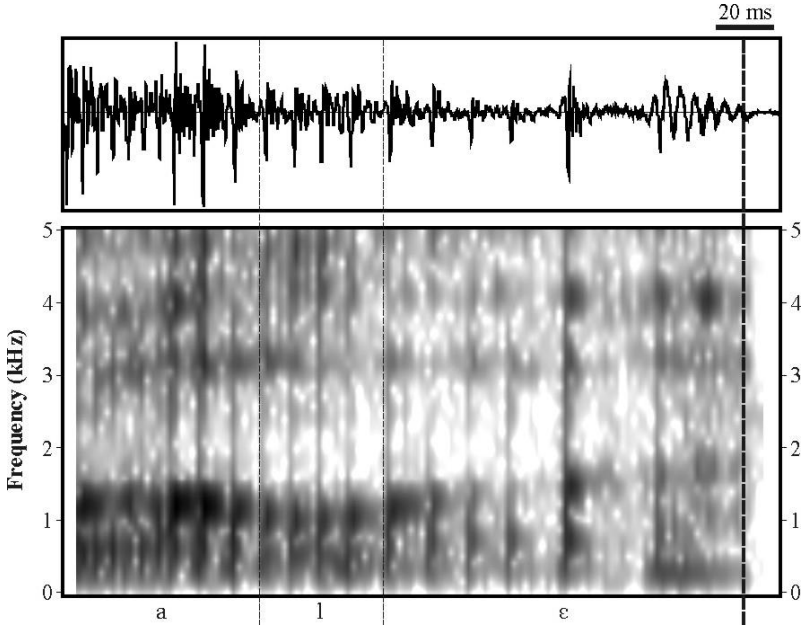


Figure 12.10. Sequence [alɛ̰̰̰] in which creaky voice is combined with breathiness in /ɛ/.

The last thing which remains to be mentioned with respect to pre-pausal phenomena is the presence of a **vocalic element**. If phonation temporally exceeds articulation, especially in sonorants, a *schwa*-like element is pronounced after the opening of the vocal tract. The vocalic element may be segmented as part of the sonorant consonant or as a separate speechsound (*cf.* Chapter 10 and Figures 10.3 and 10.6). Figure 12.11 shows the latter situation. In any case, we use the formant structure criterion for the end of the vocalic element.

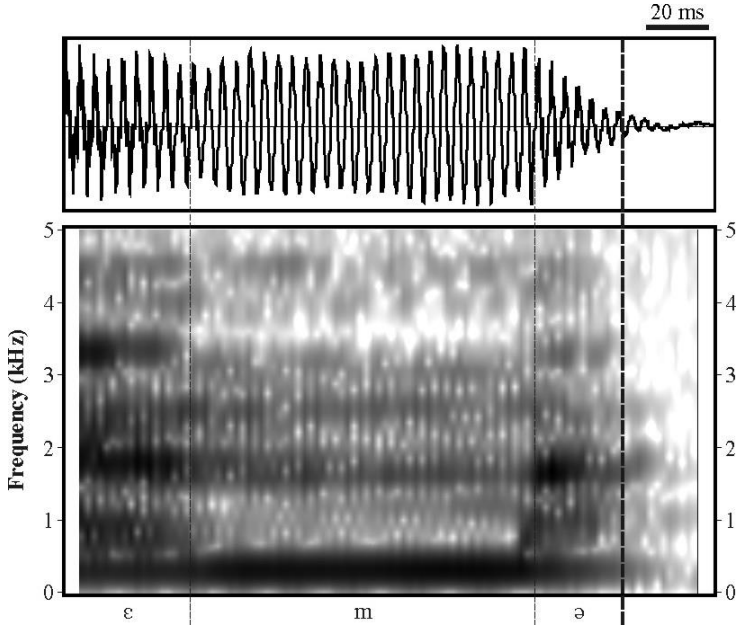


Figure 12.11. Sequence /ɛm/ at the end of the word *městem*, with an epenthetic vocalic element.

12.3. Summary

The final segmentation chapter of this handbook has dealt with some specific phenomena which are relatively frequent in utterance-initial and utterance-final contexts. They are caused by insufficient alignment between glottal and supraglottal activities.

For utterance beginnings with voiceless (or devoiced) occlusions, we recommend to allow 40-70 ms for the occlusion. Voiceless noise in initial sonorants, which is typically both visible and audible, should be included with the sonorants (see Figures 12.2 and 12.3). Preglottalization in the form of hard and soft glottal onsets (with a schwa-like vocalic element) may be segmented as independent speechsounds or as an integral part of the first "true" speechsound; this will depend, among others, on the research question at hand.

In utterance-final positions, we can encounter several ways in which phonation can cease. Listening may have to be used in instances of creaky or breathy phonation to determine whether the low-intensity sound still retains the quality of the final speechsound (typically vowel). In many utterance-final contexts, it is possible to exploit the end of formant structure as the most reliable cue for segmentation.

13. Conclusion

The objective of this book has been to propose guidelines for manual phonetic segmentation. At the heart of this handbook is our conviction that only corpora which include the target units segmented in a reliable way may be regarded as truly phonetic corpora. We believe that segmentation guidelines such as those presented in this handbook can facilitate the development of a phonetic corpus in at least two ways.

First, applying similar criteria for identifying and delimiting units in the temporal domain (individual speechsounds in this case) should result in greater comparability of the results of phonetic research.

Second, we hope to have countered, to an extent at least, the justified criticism of manual segmentation of the speech signal (see section 1.1). With rules formulated as clearly as possible, the subjective character of manual segmentation should be reduced considerably. The availability of such rules should also lead to a faster segmentation process. As a consequence, even less experienced labellers should be able to work on a corpus without compromising a uniform approach to segmentation too much.

Locating speechsound boundaries in the signal has not been treated in this book as defining the actual boundary, but as arriving at a placement which will be most sensible from the phonetic viewpoint. That is why we have tried to use inherent phonetic features of the speechsounds or speechsound classes in question to formulate our guidelines (or, more precisely, those features

which are relevant for the differentiation from neighbouring speechsounds), and why we have recommended, in cases of greater feature overlap, placing the boundary at the midpoint of a transition area so as to minimize errors.

In the Introduction, we have shown preliminary results concerning the inter-labeller reliability on a limited set of speechsound combinations. To see how these favourable results extrapolate to all combinations, the two authors labelled an approximately 90-second stretch of speech (1,360 speechsounds) taken from the Czech radio corpus after all segmentation rules have been stipulated. The mean deviation in boundary placement was 2.64 milliseconds. Recall that Pitt *et al.* (2005) report an average deviation of 16 ms and Wesenick & Kipp (1996) 10 ms.

Table 13.1 shows our deviation results in terms of increasing correct margins, so as to be able to compare them with those of Cosi *et al.* (1991, quoted in Pauws *et al.*, 1996) and Kvale & Foldvik (1991) who report 10 % and 3.5 % of boundaries differing by more than 20 ms, respectively. As the last line of Table 13.1 indicates, only 0.1 % of our boundaries (2 boundaries) differed in their placement by more than 20 ms.

correct margin	**segment boundaries**
= 0 ms	25.1 %
< 3 ms	65.6 %
< 6 ms	89.5 %
< 9 ms	97.3 %
< 15 ms	99.7 %
< 20 ms	99.9 %

Table 13.1. Correct margins in the segmentation of nearly 1,400 speechsounds, conducted by the two authors after all segmentation rules have been laid down.

It is not our purpose to analyze the discrepancies in detail, but we will mention some of the recurring problematic contexts. Of the boundary placement deviations which exceeded 9 ms (corresponding to one glottal period), the most frequent were the following:

- the trill [r], resulting from the two methods of segmenting,
- sequences of two nasals with no release of the first one,
- nasally released plosives,
- a strongly labialized [u] (with almost no energy above 1 kHz) followed by a plosive.

To conclude, we trust that the present segmentation guidelines will make the work of developing a phonetic corpus simpler, faster, and more reliable. Our attempts to increase inter-labeller reliability have indeed shown very promising results, and we sincerely hope that the same will apply for you!

References

Boersma, P. & Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1). Retrieved February 1, 2009, http://www.praat.org.

Dilley, L., Shattuck-Hufnagel, S. & Ostendiorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. Journal of Phonetics, 24, pp. 423-444.

Kiesling, S., Dilley, L. & Raymond, W. (2006). The Variation in Conversation (ViC) Project: Creation of the Buckeye Corpus of Conversational Speech, Department of Psychology, Ohio State University, Columbus, OH. Retrieved February 1, 2009, http://buckeyecorpus.osu.edu/BuckeyeCorpusmanual.pdf.

Kohler, K. J. (2007). Two Anniversaries: 75 Years of International Congresses of Phonetic Science and 50 Years of Phonetica. Some Epistemological Reflexions on the Theory and Methodology of Phonetic Science. Phonetica, 64, pp. 73-78.

Kominek, J., Bennett, C. & Black, A. W. (2003). Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis. In: Proceedings of Eurospeech 2003, pp. 313-316. Geneva: ISCA.

Kvale, K. & Foldvik, A. K. (1991). Manual segmentation and labelling of continuous speech. In: ESCA Workshop on Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication, Barcelona, paper 37.

Machač, P. (2004). Stabilita zvukových charakteristik fonémů ve spontánních mluvených projevech. In: Z. Hladká & P. Karlík (Eds.), Čeština - univerzália a specifika 5. Praha: Lidové noviny, pp. 427-435.

Machač, P. (2006). Temporální a spektrální struktura českých explozív. Unpublished PhD dissertation. Institute of Phonetics, Prague.

Machač, P. (2009). Implications of Acoustic Variation for the Segmentation of the Czech Trill /r/. In: A. Esposito & R. Vích (Eds.), Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, Lecture Notes in Artificial Intelligence 5641, pp. 173-181. Berlin, Heidelberg: Springer-Verlag.

Machač, P. & Skarnitzl, R. (2009). Phonetic analysis of parasitic speech sounds. In: Proceedings of the 19th Czech-German Workshop - Speech Processing, Prague.

Matoušek J., Skarnitzl R., Machač P. & Trmal J. (2009). Identification and Automatic Detection of Parasitic Speech Sounds. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009), pp. 876-879. Brighton: ISCA.

Pauws, S., Kamp, Y. & Willems, L. (1996). A hierarchical method of automatic speech segmentation for synthesis applications. Speech Communication, 19, pp. 207-220.

Pitt, M., Johnson, K., Hume, E., Kiesling, S. & Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Communication, 45, pp. 89-95.

Pollák, P., Volín, J. & Skarnitzl, R. (2007). HMM-Based Phonetic Segmentation in Praat Environment. Proceedings of the XIIth International Conference Speech and computer – SPECOM 2007, pp. 537-541, Moscow: MSLU.

Redi, L. & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers, Journal of Phonetics, 29, pp. 407-429.

Skarnitzl, R. (2004). Acoustic categories of nonmodal phonation in the context of the Czech conjunction "a". In: Z. Palková & J. Veroňková (Eds.), AUC Philologica 1/2004, Phonetica Pragensia X. Praha: Karolinum, pp. 57-68.

Skarnitzl, R. (2008). Koartikulační vliv nazálních konsonantů na jejich segmentální okolí v češtině a v angličtině. Unpublished PhD dissertation. Praha: FF UK.

Skarnitzl, R. (2009). Challenges in Segmenting the Czech Lateral Liquid. In: A. Esposito & R. Vích (Eds.), Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, Lecture Notes in Artificial Intelligence 5641, pp. 162-172. Berlin, Heidelberg: Springer-Verlag.

Skarnitzl, R. & Volín, J. (2005). Czech Voiced Labiodental Continuant Discrimination from Basic Acoustic Data. In: Proceedings of Interspeech 2005, the 9th Conference on Speech Communication and Technology, pp. 2921-2924. Lisbon: ISCA.

Stevens, K. (1998). Acoustic phonetics. Cambridge Massachusetts: MIT Press.

Volín, J. (2002). Čtyři scénáře vývoje české laterály. Čeština doma a ve světě, 1/2002, pp. 7-13.

Volín, J. & Skarnitzl, R. (2007). Temporal downtrends in Czech read speech. In: Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007), pp. 442-445. Antwerpen: ISCA.

Volín J., Skarnitzl R., Machač P., Janoušková J. & Veroňková J. (2008). Reliabilita a validita popisných kategorií v Pražském fonetickém korpusu. In: M. Kopřivová & M. Waclawičová (Eds.), Čeština v mluveném korpusu, pp. 249-254. Praha: Nakladatelství Lidové noviny / Ústav českého národního korpusu.

Warren, D. W., Dalston, R. M. & Mayo, R. (1993). Aerodynamics of nasalization. In: M.K.Huffman & R.A.Krakow (Eds.), Phonetics and Phonology. Volume 5. Nasals, Nasalization, and the Velum, pp. 119-146. New York: Academic Press.

Wesenick, M.-B. & Kipp, A. (1996). Estimating the quality of phonetic transcriptions and segmentations of speech signals. In: Proceedings of ICSLP 1996, pp. 129-132. Philadelphia: ISCA.

Wester, M., Kessens, J. M., Cucchiarini, C. & Strik, H. (2001). Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer. Language and Speech, 44, pp. 377-403.