Tutorial:

Principles of Phonetic Segmentation

Xiaoye Wu

Feburary, 2025

Motto – Everything has boundaries, though often unclear.

This is a brief tutorial based on the book *Principles of Phonetic Segmentation* written by Pavel Machač and Radek Skarnitzl in 2009. It is organised based on different combinations of segments, providing quick guidelines for phoneticians to place boundaries to distinguish neighbouring sounds.

1 General ideas

- **Why do we need phonetic boundaries?**

  In order to understand the structure and function of our continuous speech, we have to find a consistent and reliable way to divide the continuous acoustic signal into smaller chunks for closer investigations.

For different research purposes (e.g., segmental properties or prosodic studies), we may have different scopes and standards for segmentation.

We have two ways to realise segmentation: manual or automatic, which both have its pros and cons. It is usually a battle between accuracy and efficiency. As every phonetician may have such experiences, manual segmentation is extremely time-consuming (especially when you have hours and hours of audio recordings). Subject and varying standards may be applied each time when boundaries are laid down. Both intra-labeller and inter-labeller consistency are at risk. Automatic segmentation can save some energy as suitable pre-trained models are selected. However, manual check and corrections are still needed to ensure accurate segmentations adapted to your own data. Thus, it is significant to propose guidelines for segmentation to enhance accuracy, no matter what methods you use.

- **What do we mean by 'the boundary'?**

  Bear in mind, there is no 'true' boundaries existing between segments in a continuous flow. The boundaries we mark are tools used to facilitate phonetic research.

  In ideal cases, the contrasts between segments are salient, so the boundary location is quite straightforward. However, in most real-life cases, the transition phases between segments are long and some phonetic

properties of the neighbouring segments may overlap, thus marking the boundary location is ambiguous.

The guidelines aim to increase the reliability and unification of segmentation by stipulating relatively simple and unambiguous rules which are mainly based on inherent phonetic features.

- **What are inherent and extrinsic phonetic features?**

  The enumeration of inherent phonetic features constructs a given speech-sound in its full and canonical form. Inherent features can be viewed from different perspectives: acoustic, articulatory and perceptual. For instance, from the acoustic viewpoint, vowels feature the presence of fundamental frequency and formant structure. From the articulatory viewpoint, vowels are characterised by the opening of the vocal tract and the vibration of the vocal folds. From the perceptual viewpoint, vowels include high sonority, voice quality and so on. Among these features, some are stable such as formant structure, while some are less stable such as voice quality. See Figure 1.1 for some stable and less stable inherent phonetic features of different kinds of sounds in Czech.

  Comparatively, extrinsic phonetic features are dynamic and highly variable, which is caused by coarticulation, for example. Some phonetic features extend beyond the boundary of the segment and spread to its neighbouring segments. It is these extrinsic features that makes

segmentation more difficult and ambiguous.

Therefore, we are employing inherent phonetic features as the references for segmentation.

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	

Figure 1.1: Stable and less stable inherent phonetic features in Czech

- **What are the general principles of segmentation?**

  1. The Full Formant Structure Rule: Formant structure (a very useful and repeatedly used feature) is indicated by salient formant columns (i.e., the dark vertical bars in the spectrogram, see Figure 1.2). Boundaries are placed next to or between the bars.

  2. The Midpoint Rule: if there is a transition phase, the boundary will be places at the temporal midpoint of this phase.

  3. The Zero Crossing Rule: All boundaries will be places at a zero crossing (see Figure 1.2).

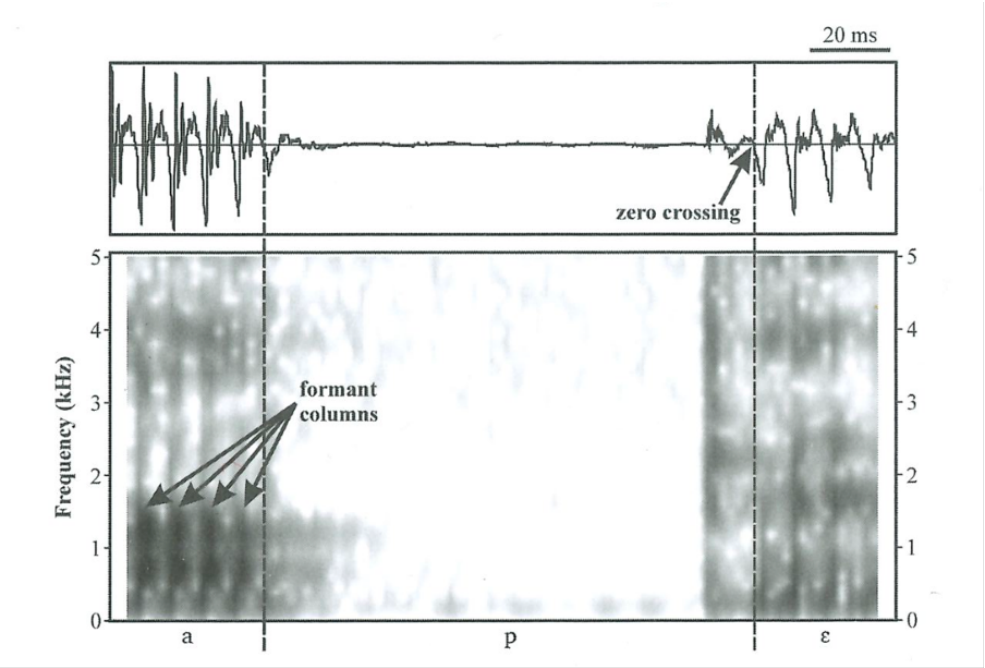

Figure 1.2: Stable and less stable inherent phonetic features in Czech

- **What is the recommended pipeline for segmentation?**

  I organise the steps of placing boundaries in different contexts and summarise it as a general pipeline of realizing segmentation.

  Step 1: Choose the related inherent phonetic feature(s)

  Step 2: If the contrast between neighbouring sounds is salient, identify the boundary in the spectrogram or in the waveform.

  Step 3: If the contrast between neighbouring sounds is not salient and there is a transition phase, place the boundary at the midpoint.

  Step 4: If unfortunately, neither the spectrogram nor waveform works, place the boundary based on the hearing judgment.

Tips:

Just identify the differences between neighbouring elements!

Make good use of both visual and auditory cues!

- **Tailor to your own data!**

  Keep in mind that the principles proposed in this book is mainly built upon Czech and English data, which might not be universal and serve as the best principles for the language you are working on. Be careful when using them and tailor the rules to your own data if needed. Also, there may be some segments and combinations of segments which are not mentioned in this tutorial, propose your own rules accordingly.

Different combinations of sounds will be included and guidelines for placing the boundaries will be provided respectively. It will comprise intervocalic consonants (including plosives, fricatives, nasals, trills, glides, and lateral alveolar approximants), consonant clusters (including obstruent clusters, obstruent-liquid sequences, sequences of sounds with the same manner of articulation), glottal stop in word-initial vowels, and utterance-initial and utterance-final contexts. In each section, relevant phonetic features will be introduced first, followed by the proposed segmentation rules and additional guidelines for some less straightforward cases.

For quick look-ups, I organised the rules for each category of segment

combinations as below. For some less straightforward cases, please refer to the section 'Additional segmentation guidelines' in each chapter.

Segments	Principles
Intervocalic plosives	Full formant structure
Intervocalic fricatives	Full formant structure; Relative intensity differences
Intervocalic nasal consonants	Low intensity in high-frequency intensity
Intervocalic trills	The 'cycle-oriented' approach: low-intensity cycle The 'extended' approach: cycle + surrounding low-intensity vocalic part
Intervocalic glides	The acoustic approach: midpoints of the formant transitions The perceptual approach: auditory cues
Intervocalic lateral alveolar approximants	Low intensity in high frequencies Relative formant intensity (weaker F2 and stronger F3) Simpler waveform
Obstruent clusters (fricative/affricate+plosive)	Full fricative noise
Obstruent-liquid sequences	Full formant structure
Clusters of two plosives/nasals	Closure-release sequences
Clusters of two fricatives	Different intensity in high frequencies
Glottal stops	Plosive-like: Closure-release sequences Creaky: Aperiodicity
Utterance-initial and utterance-final contexts	Initial: 40-70 ms closure interval Final: Full formant structure

2 Intervocalic plosives

2.1 Acoustic and articulatory features

Articulatory: Two stages–closure and release

Acoustic:

Voiceless–No fundamental frequency nor formant structure in the closure phase; burst release is short and salient

Voiced–F0 throughout the closure phase; release is weak and usually without aperiodic noise

2.2 Inherent phonetic features

- Complete closure and the consequent absence of formant structure
- Presence of release
- Oral character
- Presence of F0 in phonologically voiced plosives and absence of F0 in phonologically voiceless plosives

2.3 Features for segmentation of intervocalic plosives

Primary cue: Full formant structure

Onset: Absence of formant structure

Offset: Presence of formant structure

See Figure 2.1 which represents the continuous acoustic events of [ucɪ]. Stage (2)-(5) are regarded as the plosive [c], with no formant structure. (2) Voicing continuation and (5) F0 onset are part of the plosive as well.

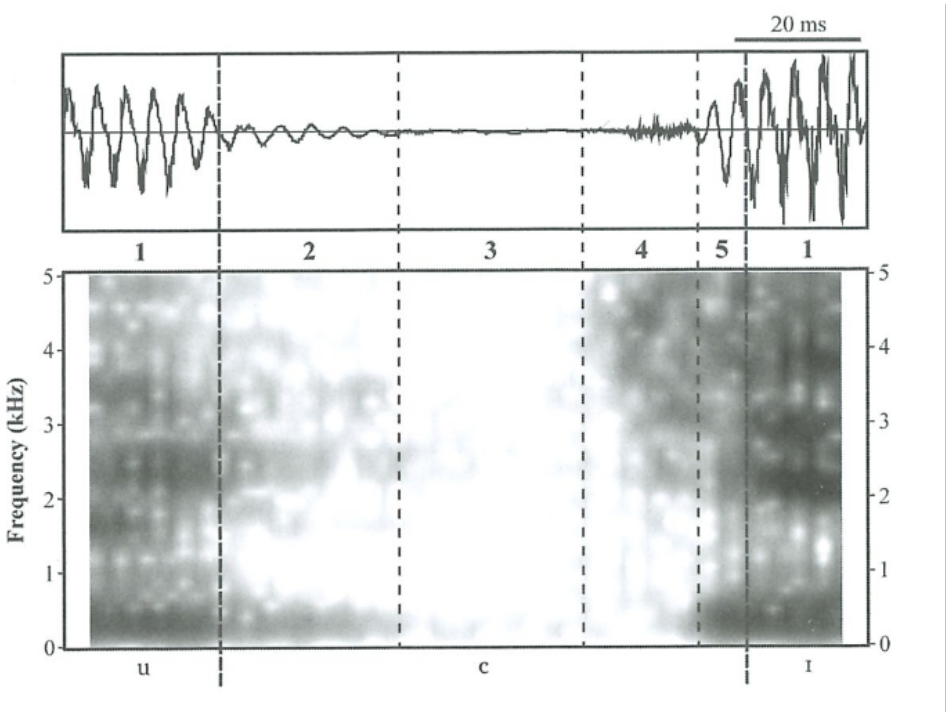

Figure 2.1: Sequence [ucɪ]: (1) vowel; (2) voicing continuation into the closure; (3) complete closure without voicing; (4) release in the form of noise burst; (5) F0 onset of following vowel.

2.4 Additional guidelines

- Spirantisation of the plosive leads to the absence of complete closure – stick to the Full Formant Structure Rule

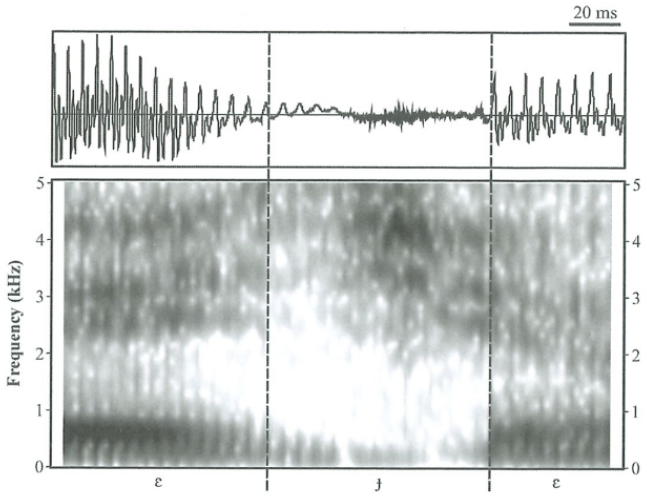

Figure 2.2: Sequence [ɛɟɛ] with strong spirantisation

- Multiple releases of velar plosives – The Full Formant Structure Rule

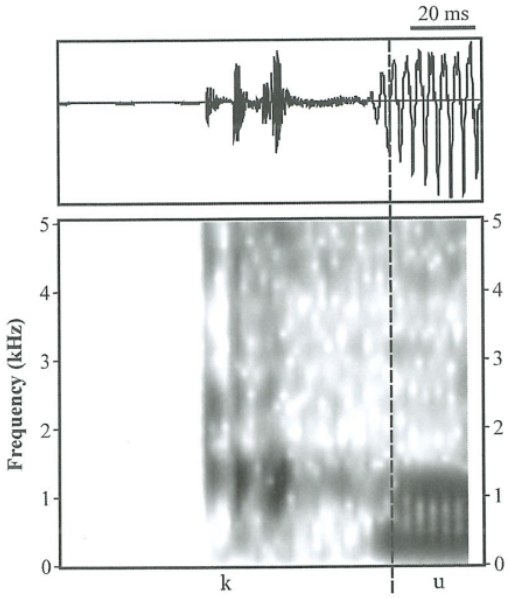

Figure 2.3: Sequence [ku] with multiple releases

- Long transition phases – The Midpoint Rule

  Several forms of the transition phases:

  The gradual decay of the formant structure;

  Formant columns are not salient;

  A noise component interferes the formant structure.

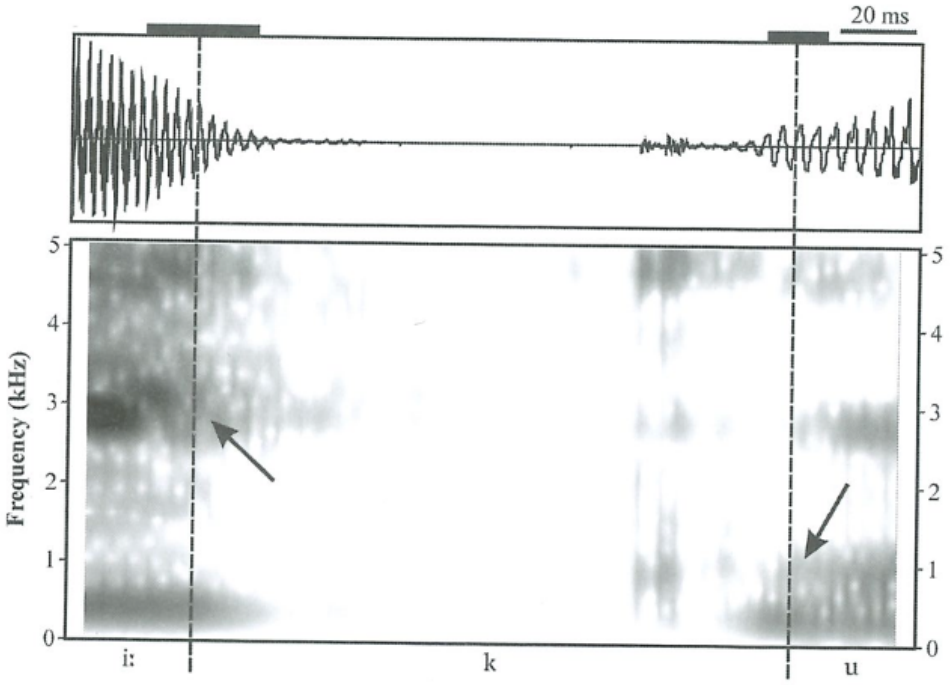

Figure 2.4: Sequence [i:ku] with long transition phases consisting slow decay of formant structure (The upper horizontal black bars indicate the extent of formant decay and onset and the arrows indicate the mid-point boundaries.)

- Palatal noise – Place the boundary at the midpoint of the noise section

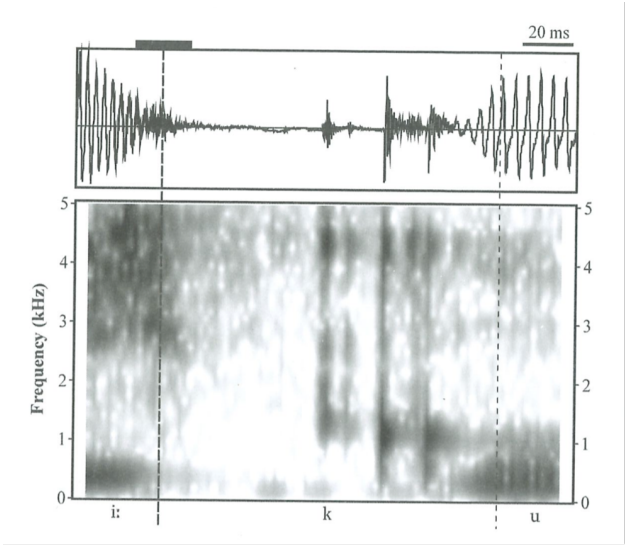

Figure 2.5: Sequence [i:ku] with palatal noise

- Flap realisation of the alveolar /d/ – The Full Formant Structure Rule / Use auditory cues

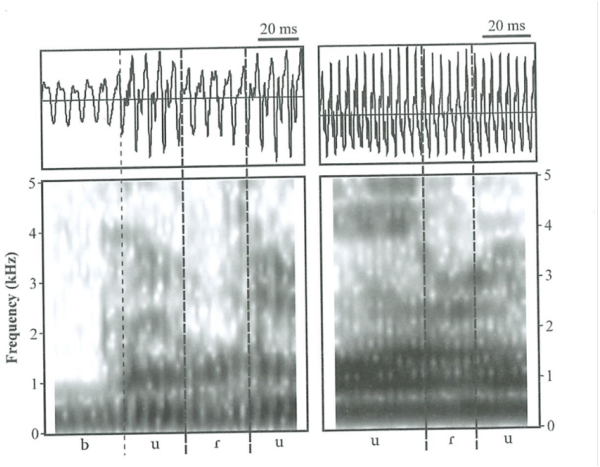

Figure 2.6: Czech word *budu* with flap realisation: weak formant structure in high frequencies (left); no visual cues (right)

3 Intervocalic fricatives

3.1 Acoustic and articulatory features

Articulatory: stricture – narrowing of the airstream passage generates turbulence and creates fricative noise

Acoustic:

Voiceless–No F0; noise formant (noise with only aperiodic component)

Voiced–Both noise and periodic components

3.2 Inherent phonetic features

- Presence of noise components
- Presence of fundamental frequency in phonologically voiced fricatives and absence of fundamental frequency in phonologically voiceless fricatives

3.3 Features for segmentation of intervocalic fricatives

- Primary cue: Full formant structure

  Onset: Absence of formant structure

  Offset: Presence of formant structure

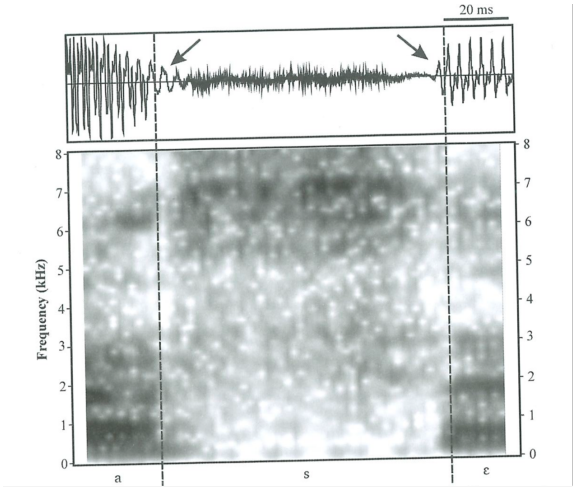

Figure 3.1: Sequence [asɛ] showing the segmentation of the intervocalic voiceless fricative [s] based on vowels' full formant structure

- Secondary cue: Relative intensity differences

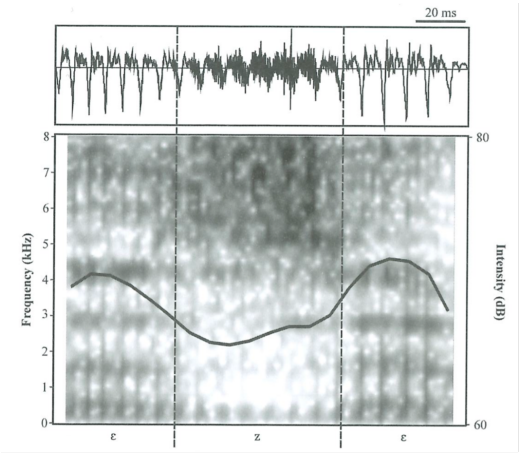

Figure 3.2: Sequence [ɛzɛ] showing the segmentation of intervocalic voiced fricative [z] where the boundary is placed around the midpoint of the intensity drop or increase

3.4 Additional guidelines

- Slow decay or onset of the formant structure – The Midpoint Rule

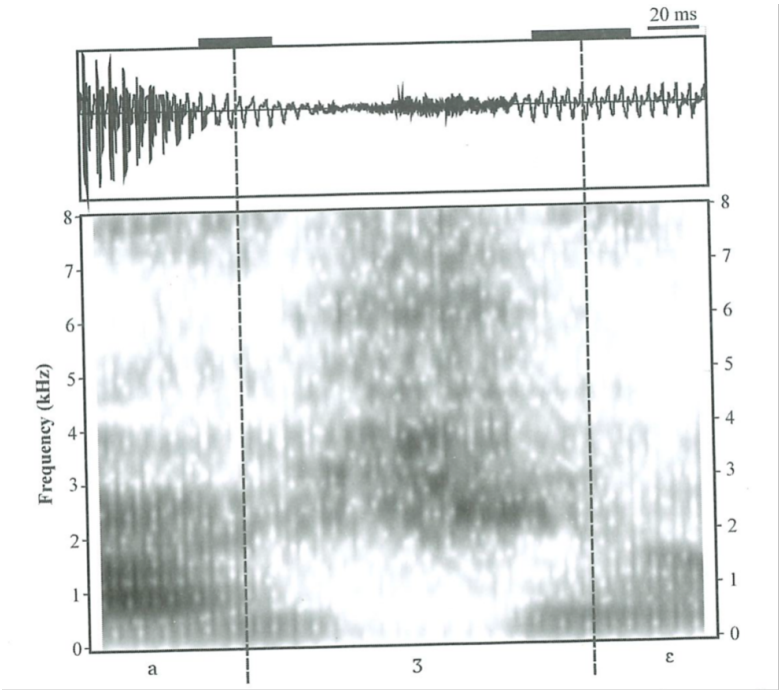

Figure 3.3: Sequence [aʒɛ] showing the long transition area in vowel-fricative and fricative-vowel sequences where the boundary is placed at the midpoint of the transition phase

- Approximantisation/Lenition of the intervocalic /v/ - differences in formant intensity

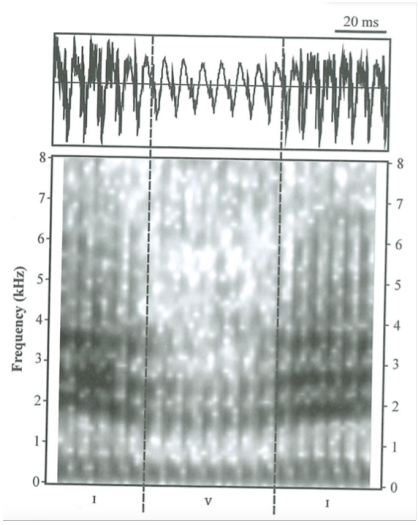

Figure 3.4: Sequence [ɪvɪ] from the English word *giving* showing salient differences in formant intensity (especially around 4 kHz)

- Voiced laryngeal fricative [ɦ] – broader formant bandwidths and 'spiky' waveform

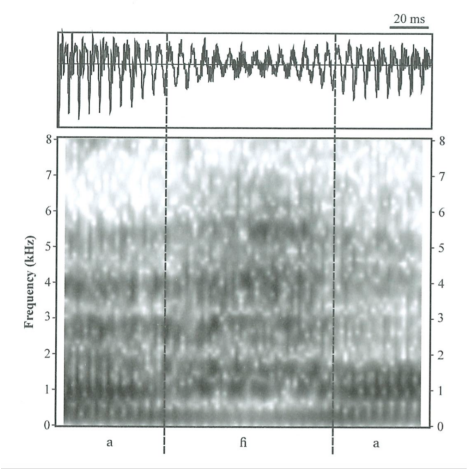

Figure 3.5: Sequence [aɦa] with salient noise and broader formant bandwidths

- Segmenting affricates: Left boundary identical to plosives; Right boundary identical to fricatives

4 Intervocalic nasal consonants

4.1 Acoustic and articulatory features

Articulatory: Lowered velum; Closure in the oral cavity

Acoustic:

Presence of nasal formants (usually N1 around 250 kHz, N2 slightly below 1000 kHz, and remain the same for all nasals in a given speaker); Presence of the oral antiformants (A1 differs according to places of articulation, lower the amplitude of spectrum above the antiformant

4.2 Inherent phonetic features

- Voicing and high sonority
- Velum lowering
- Closure in the oral cavity – presence of antiformants –energy concentrated in lower frequencies

4.3 Features for segmentation of intervocalic nasals

- Primary cue: Low intensity in high-frequencies ('simpler' shape of the waveform-no spiky or hairy waveform)

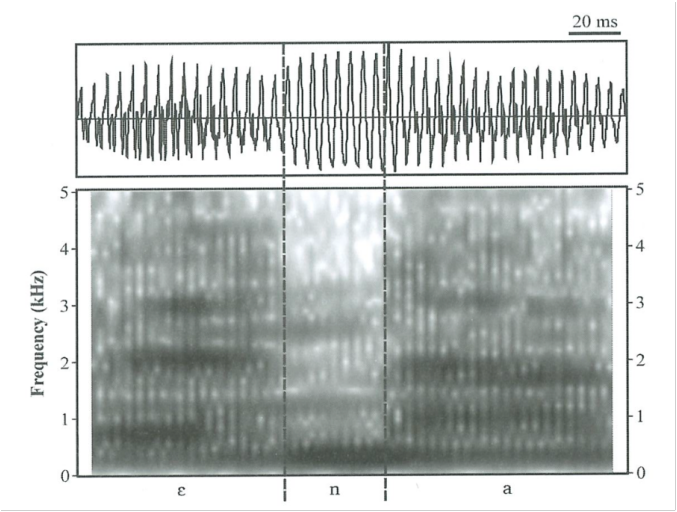

Figure 4.1: Sequence of [ɛna] showing nasal consonant [n] with low intensity in high-frequencies and the simpler shape of waveform (Sometimes also the presence of antiformant indicated by the light horizontal stripe and visible formant transitions)

Generally, we can use one or multiple cues to determine the boundaries for nasals. If one does not work for your data, then move on to the next rule. Multiple rules can be used together to validate the accuracy of the boundaries. If none of the visual cues works, use listening to facilitate.

Onset: Beginning of the lower-amplitude period in the waveform/Beginning of the lower-intensity in high-frequencies in the spectrogram/Absence of vowel formants/Presence of antiformants

Offset: Beginning of more salient vowel formants/Beginning of the more 'complex' waveform/Presence of vowel transitions/End of plosive-like elements with a higher amplitude

5 Intervocalic trills

5.1 Acoustic and articulatory features

Articulatory: Fast and repeated forming and releasing of a stricture

Acoustic: Salient formant structure (F1 450Hz, F2: 1300-1400 Hz, F3 2000 Hz); Alternating vocalic part and the decay of acoustic energy

5.2 Inherent phonetic features

- Presence of at least one period (vocalic part – cycle – vocalic part)
- Presence of voicing

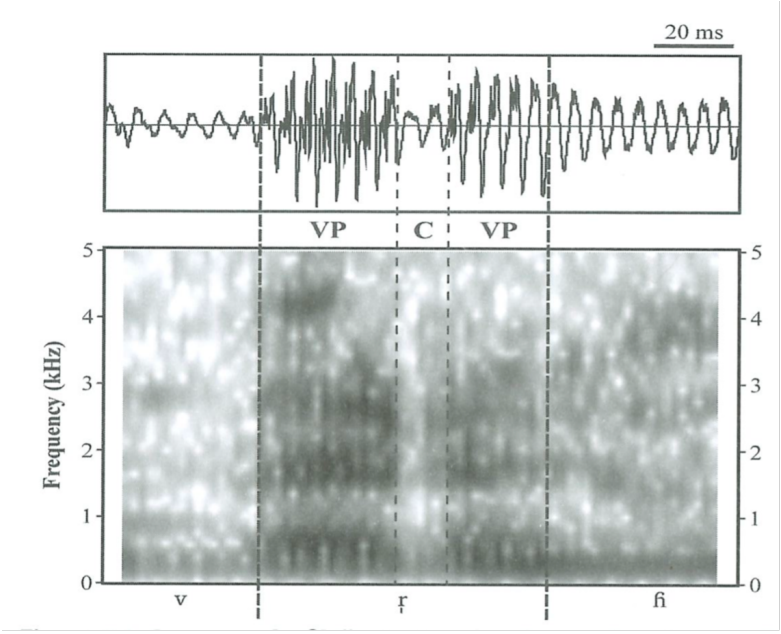

Figure 5.1: Sequence [vr̥fi] showing the period of the trill [r̥] with two vocalic parts (VP) and one low-intensity cycle (C)

However, when the trill is intervocalic, the acoustic contrast between the vocalic part of the trill and the neighbouring vowels is usually very low, posing problems for segmentation (as shown in Figure 5.2).

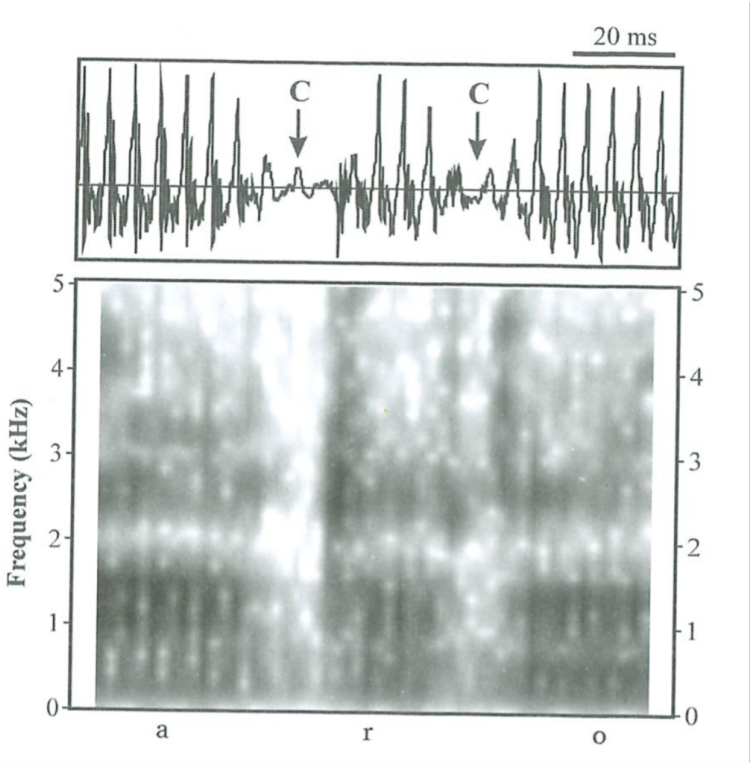

Figure 5.2: Sequence [aro] with two salient cycles indicated by arrows, but the vocalic parts are difficult to be distinguished by visual cues

There are two ways in which we can deal with the vocalic parts: the 'cycle-oriented' way and the 'extended' way:

5.2.1 The 'cycle-oriented' way

Only the cycle itself constitutes the trill.

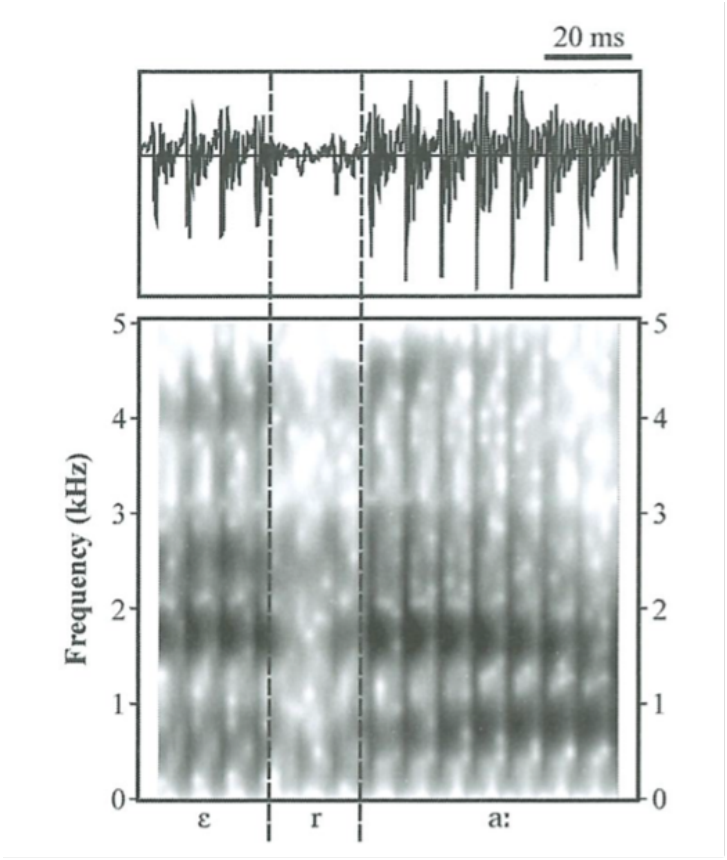

Figure 5.3: Sequence [ɛra:] showing the segmentation where only the cycle constitutes [r]

5.2.2 The 'extended' way

In some cases, the vocalic parts are visible, thus we should include the transition phases into the trill.

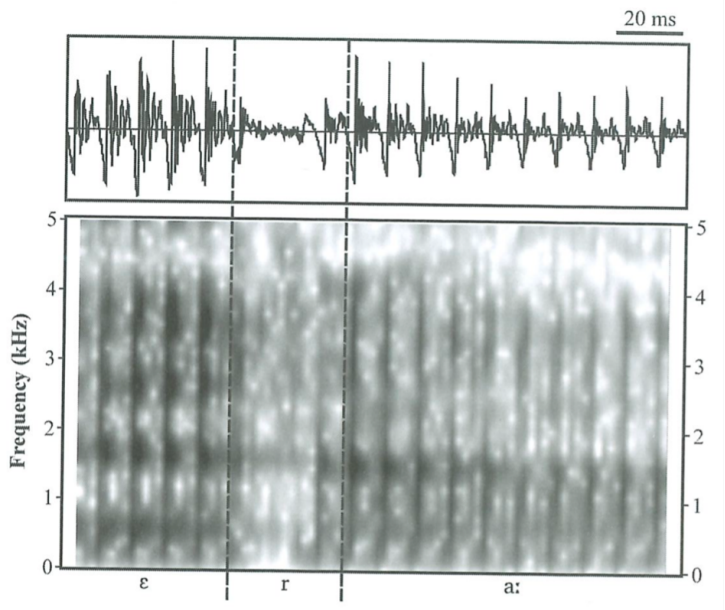

Figure 5.4: Sequence [ɛra:] showing the periods around the cycle identifiable from the vowels with salient lower amplitude in the waveform and weaker intensity in the spectrogram (usually one period before and after the cycle)

6 Intervocalic glides

6.1 Acoustic and articulatory features

Articulatory: Approximation of the active articulator to the passive organ with no turbulence in friction

Acoustic:

Palatal [j] and labio-velar [w] with similar formant values to their vowel counterpart [i] and [u];

Post-alveolar approximant [ɹ] (BrE) or retroflex [ɻ] (AmE) with very low F3 (1.5-2 kHz)

6.2 Inherent phonetic features

- Full formant structure
- Presence of voicing
- Higher F2 in [j] (convex shape), Lower F2 in [w] (concave shape), Lower F3 in [ɹ] (concave shape)

Since the inherent phonetic features of glides are identical to those of vowels, the acoustic contrast between the glide and its surrounding vowels is quite low. There are two approaches of segmentation: the acoustic approach and the perceptual approach.

6.2.1 The acoustic approach – The Midpoint Rule

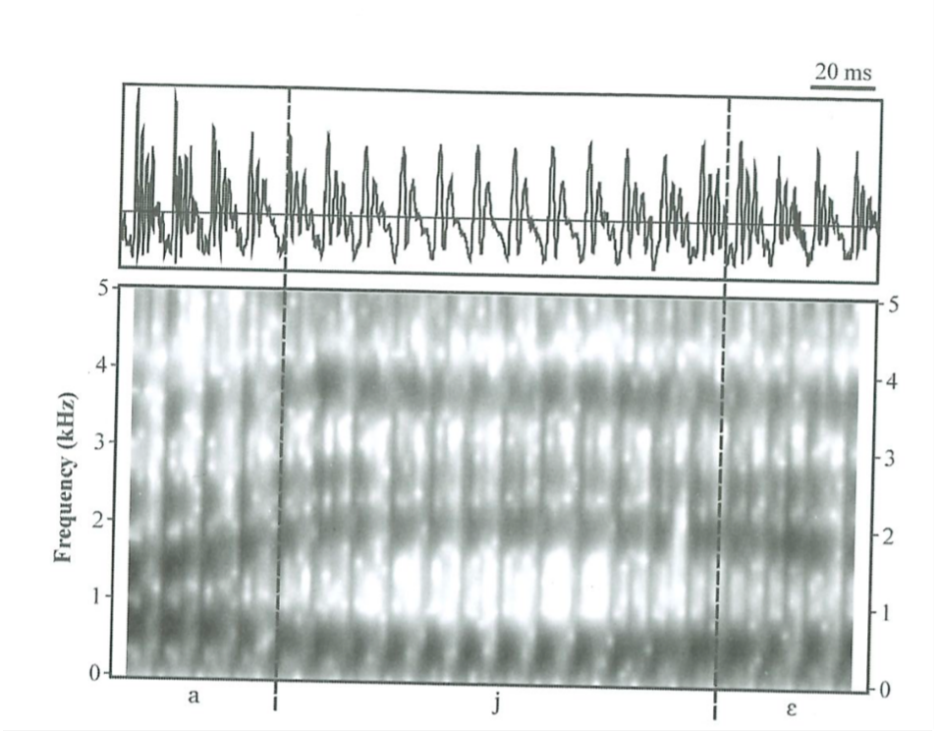

Figure 6.1: Sequence [ajɛ] showing the segmentation of the intervocalic [j] with boundaries placed at the midpoints of F2 transitions

The advantage of taking the acoustic approach is that it applies a uniform segmentation rule based on visual cues. However, it leads to a false auditory impression and a longer duration than the perceptual approach, failing to reflect the key perceptual feature of the glides, namely their non-syllabic characteristics.

6.2.2 The perceptual approach – Use auditory cues

Take sequence [oja] as an example, we listen and try to find the moment when hearing /oj/ and /ja/ as monosyllables rather than two syllables. We put the right boundary when we hear a monosyllabic sequence of [oj] rather than [oj^{ə}] with a vocalic tail. Similarly, we put the left boundary when we hear a monosyllabic sequence of [ja] rather than [^{ə}ja] with a vocalic head.

The perceptual approach can apply to all cases where there are no visual cues, but it is very time-consuming and also subjective.

6.3 Additional segmentation guidelines

- 'Fortis' pronunciation of /j/: treat as a voiced palatal fricative [ʝ]
- Relative explicit but not fricated /j/: midpoint of the amplitude envelope changes

7 Intervocalic lateral alveolar approximants

7.1 Acoustic and articulatory features

Articulatory: A closure along the mid-sagittal line of the oral cavity with the sides of the tongue lowered

Acoustic (For the canonical alveolar /l/):

Formants structure: F1 – 350 Hz, F2 – 1300 Hz, F3 – 2800 Hz;

Antiformant: between 2-3 kHz

NOTE Laterals frequently undergo coarticulation and have high acoustic variability, so they are quite challenging in terms of segmentation.

7.2 Inherent phonetic features of alveolar laterals

- Voicing and high sonority
- A closure in the alveolar region
- Presence of an antiformant

7.3 Features for segmentation of intervocalic laterals

The cues for segmentation of laterals are identical to those of the nasals.

- Presence of an antiformant around 2-3 kHz
- Low intensity in high-frequency region (F4 and F5 region)

- 'Simpler' shape of waveform

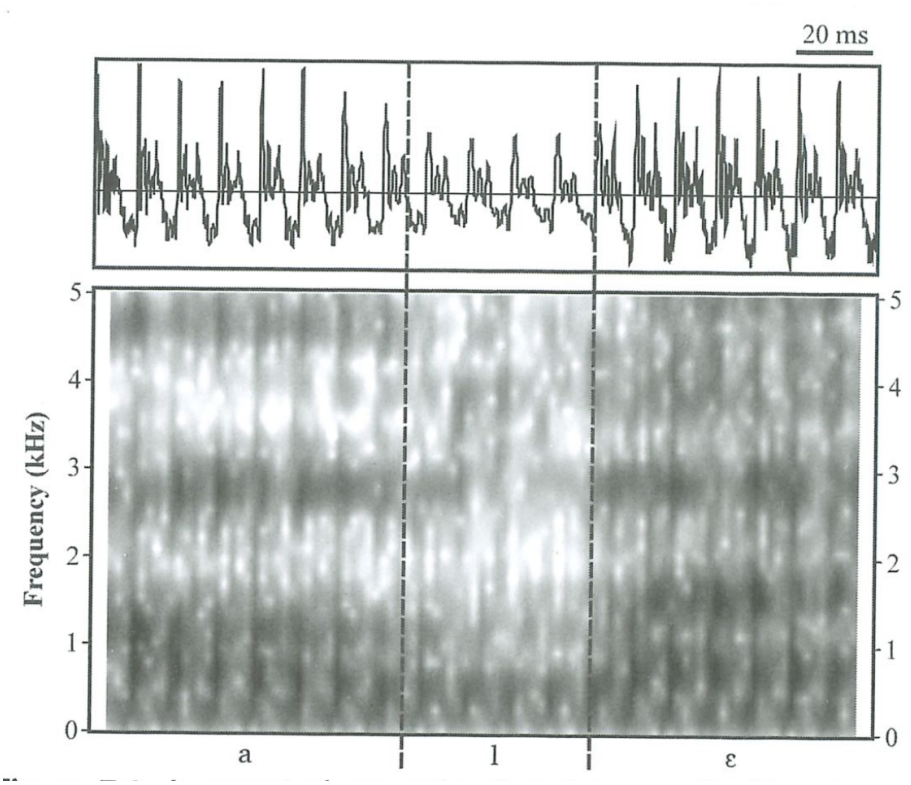

Figure 7.1: Sequence [alɛ] showing all three cues available for segmentation

As /l/ represents high variability, these three cues are not always visible. More guidelines are provided for those less canonical laterals.

7.4 Additional segmentation guidelines

Aside from the three cues mentioned above, relative intensity of formants can also serve as effective cues in some cases. For instance, /l/ shows more salient F3 (shown in Figure 7.2) and weaker F2 (shown in Figure 7.3).

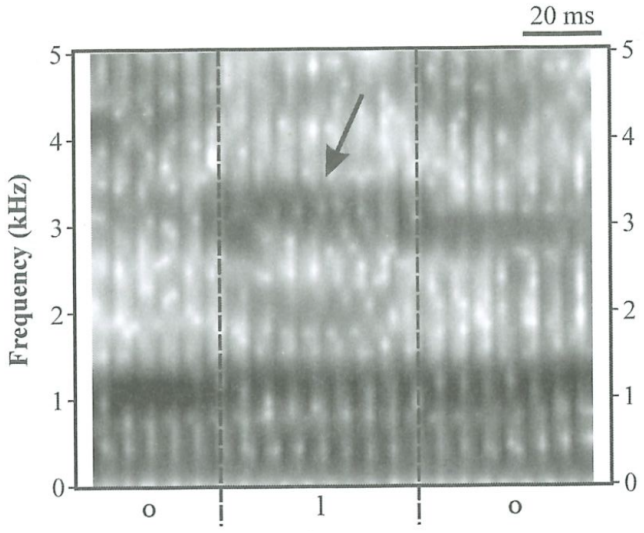

Figure 7.2: Sequence [olo] showing more salient F3 of /l/ than its neighbouring vowels

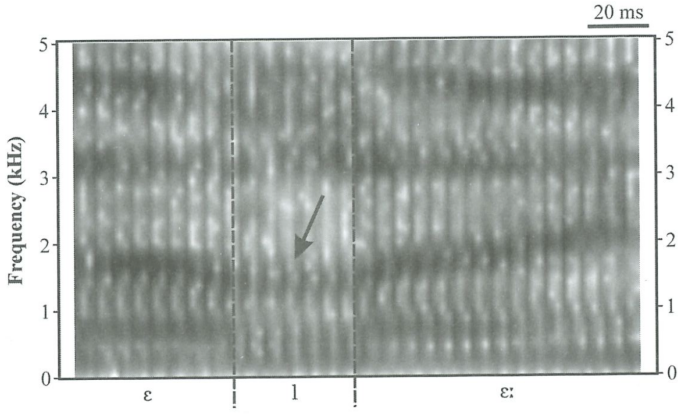

Figure 7.3: Sequence [ɛlɛː] showing weaker F2 of /l/ than its neighbouring vowels

There are many cases of /l/ requiring auditory cues, especially when /l/ is vocalised.

8 Obstruent clusters of different manner of articulation

This chapter mainly deals with clusters of fricatives/affricates and plosives.

8.1 Basic segmentation rules

Primary cue: Full fricative noise

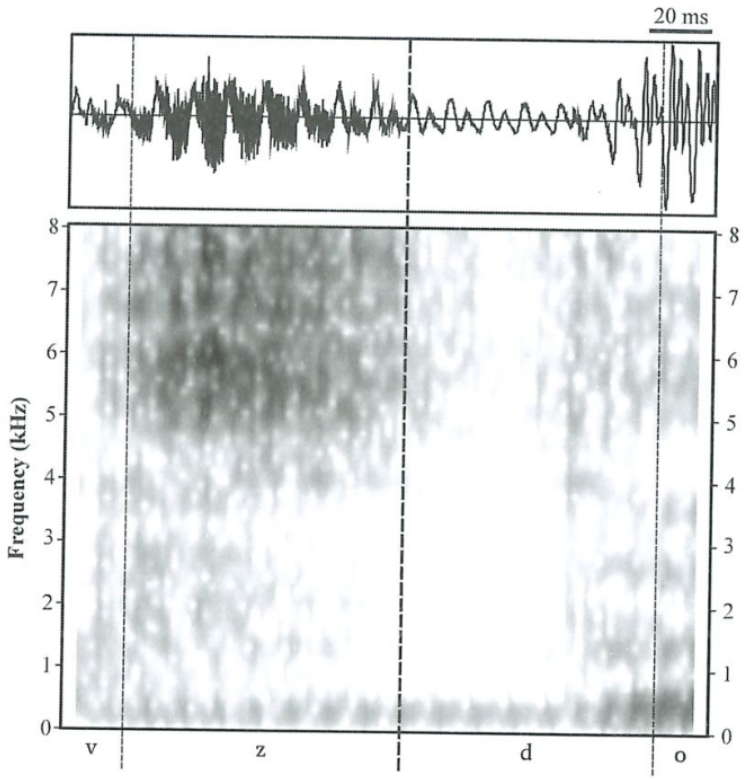

Figure 8.1: Sequence [vzdo] showing clear noise in fricative [z] and no noise in [d]

8.2 Additional segmentation guidelines

- Residual noise – Considered as part of the following plosive

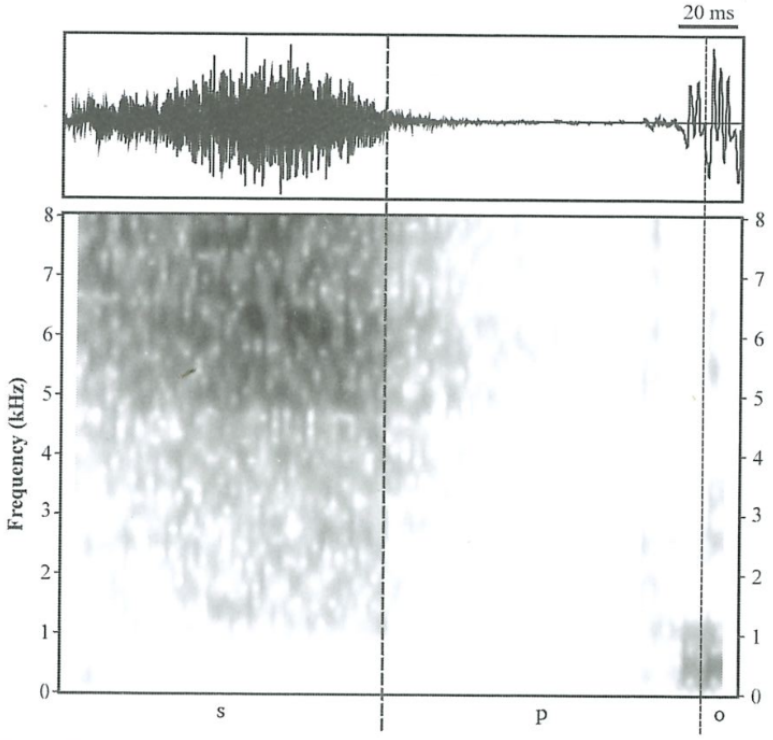

Figure 8.2: Sequence [spo] showing the boundary placed at the end of the friction and the residual noise belonging to the closure phase of the plosive [p]

- Gradual decay of noise – The Midpoint rule

  The transition phase is marked by the beginning/end of the full-fledged noise and the end/beginning of the residual noise

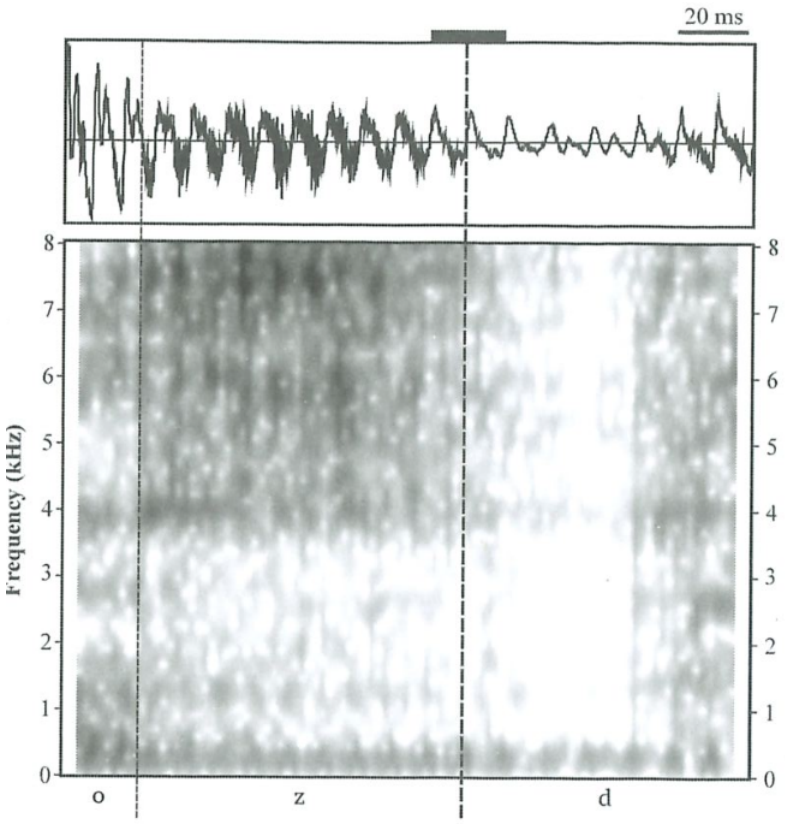

Figure 8.3: Sequence [ozd] showing the boundary placed at the mispoint of the transition phase

9 Obstruent-liquid sequences

The main segmentation rule regarding the liquid [l] is based on the full formant structure.

- Devoicing of /l/ in [sl̥]

  As shown in Figure 9.1, intensity of high-frequency noise in [s] is higher than that in [l̥].

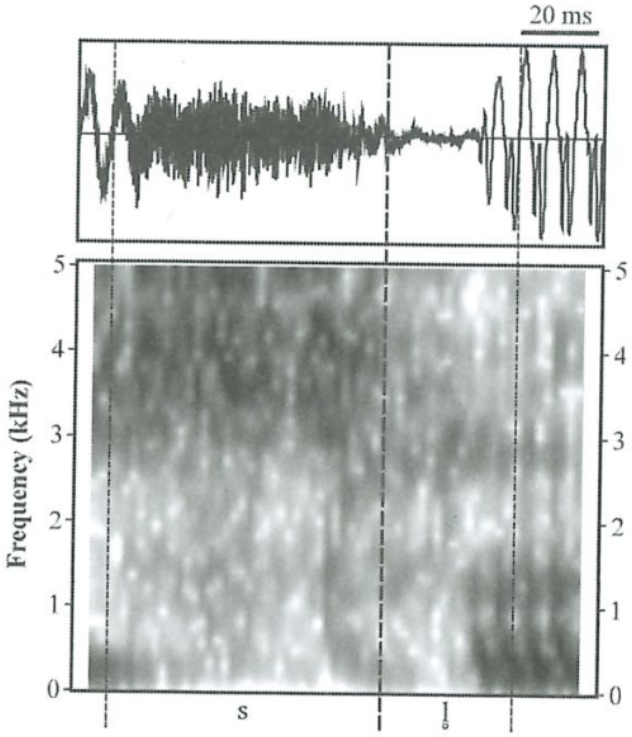

Figure 9.1: Sequence [sl̥] showing more noise in high frequencies in [s]

- Epenthetic [t]-like element due to the overlap of articulatory gestures in /sl/ [s^{t}l̥]

As represented in Figure 9.2, the epenthetic element is regarded as part of the fricative [s].

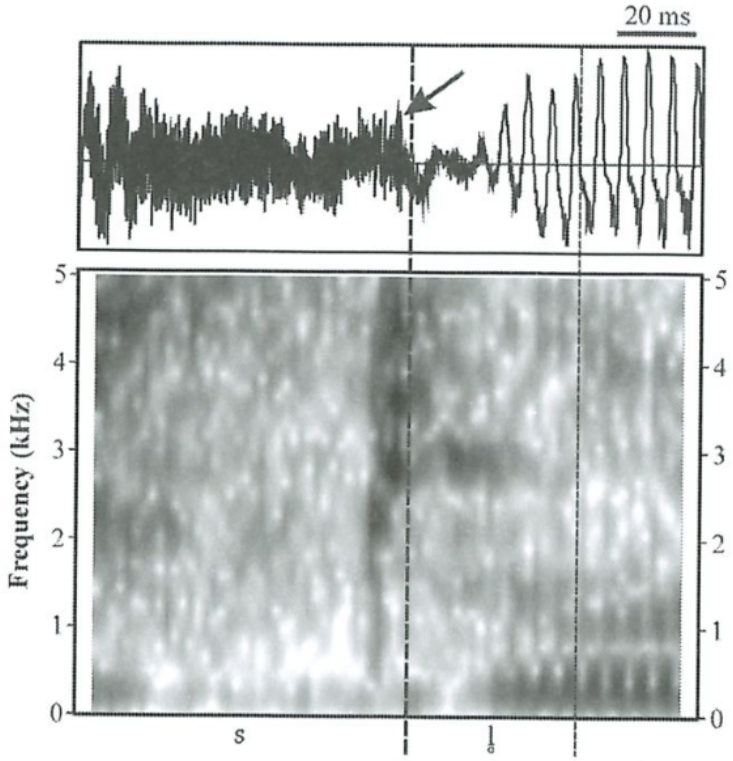

Figure 9.2: Sequence [s^{t}l̥] showing the epenthetic [t] (indicated by the arrow) as part of the fricative [s]

- Partial devoicing and lateral release of plosive in /tl/ [tl̥]

  Shown in Figure 9.3, the boundary is placed at the onset of the formant structure of the voiced part of /l/.

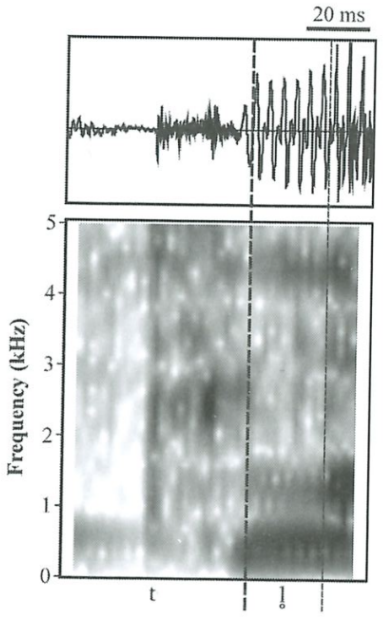

Figure 9.3: Sequence [tl̥] showing the lateral noise as part of the plosive [t]

10 Sequences of speechsounds with the same manner of articulation

10.1 Clusters of two plosives and nasals

- Succession of two plosives – two closure-release sequences (with either complete or incomplete closure)

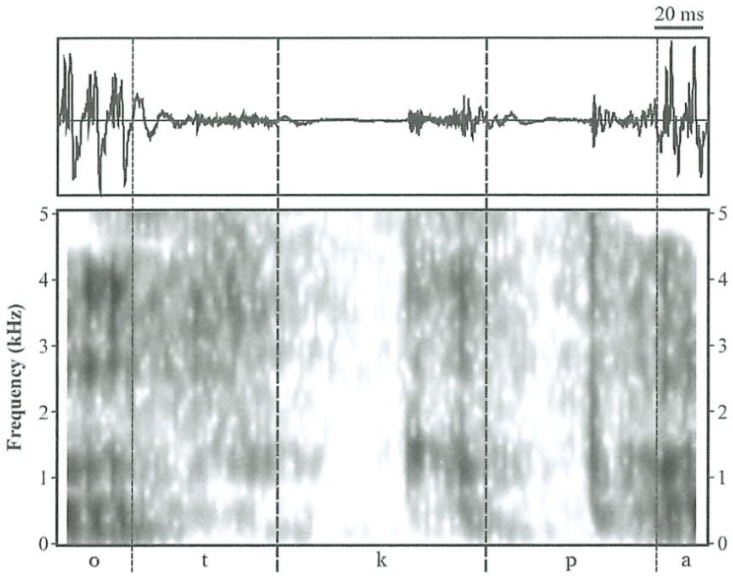

Figure 10.1: Sequence [otkpa] with incomplete closure in [t] and fully released [k] and [p]

- Epenthetic schwa-like element – fortification of the first plosive – segmented as part of the first plosive or as an independent element

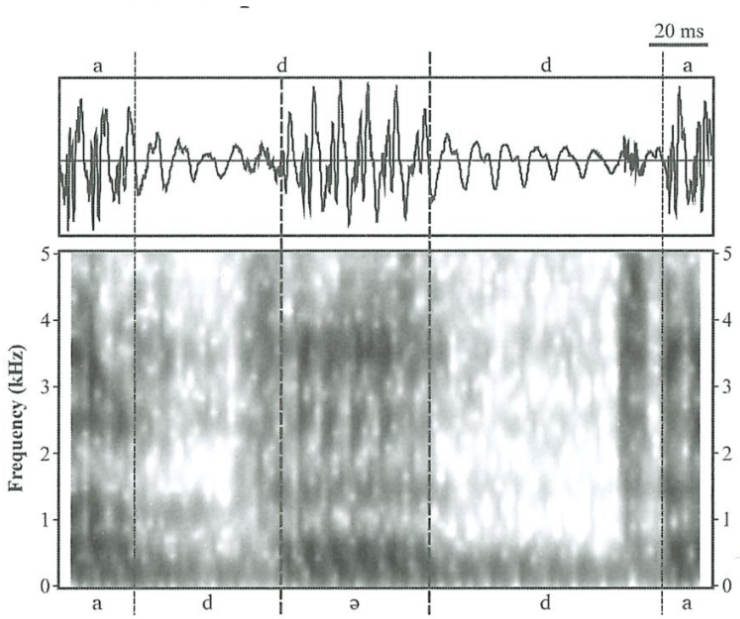

Figure 10.2: Sequence [adəda] with schwa segmented as part of the first plosive or as an independent segment

- Unreleased first plosive – No release phase – Place the boundary at the midpoint of the long double closure phase

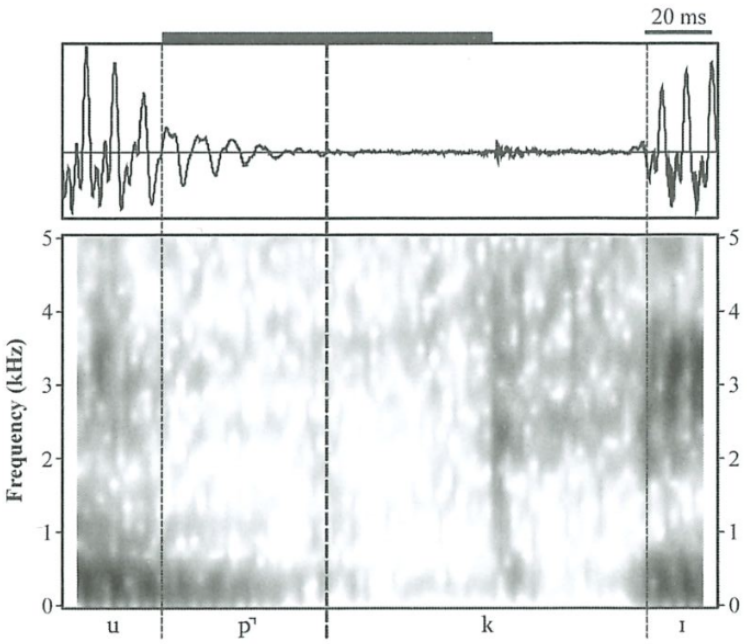

Figure 10.3: Sequence [upkˀɪ] showing the boundary placed at the midpoint of the long closure phase

The segmentation rules regarding two consecutive nasals are identical to plosives.

10.2 Clusters of two fricatives

- If the two fricatives are of different place of articulation, spectral properties (noise formants and relative intensity) usually show differences which can serve as cues for segmentation.

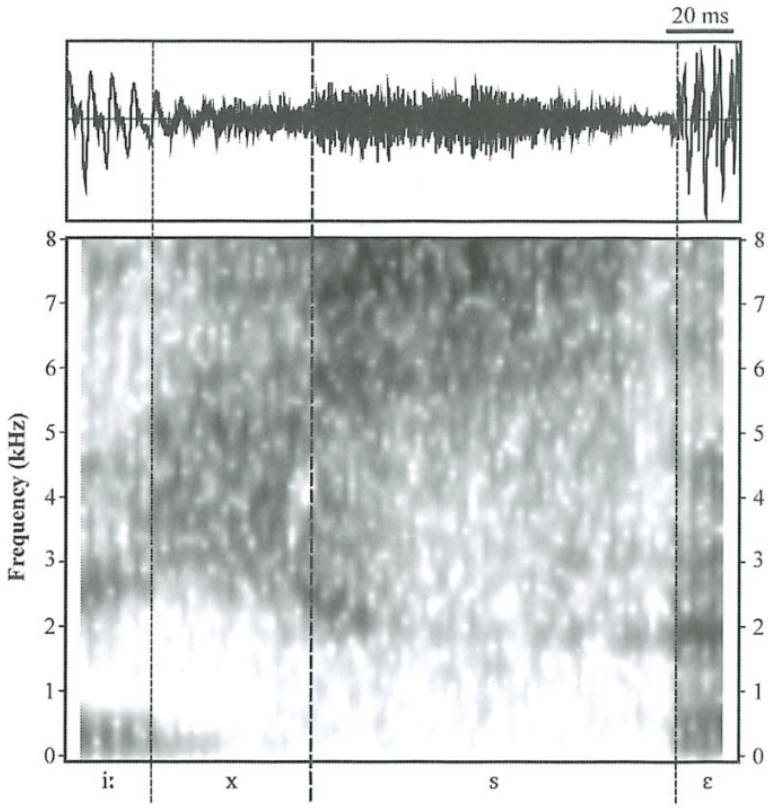

Figure 10.4: Sequence [i:xsɛ] showing the difference of intensity in high frequencies

- If the two fricatives are of the same place of articulation, segmentation is usually difficult which requires auditory facilitation.

11 The glottal stop in word-initial vowels

There are usually two types of glottal stop preceding vowels: plosive-like or creaky.

11.1 Plosive-like glottal stop

A canonical glottal stop consists of the closure phase and the release, so we can easily treat it as a plosive.

- Right boundary: Usually contains formants of the next vowel – treat high-intensity periods as part of the glottal stop

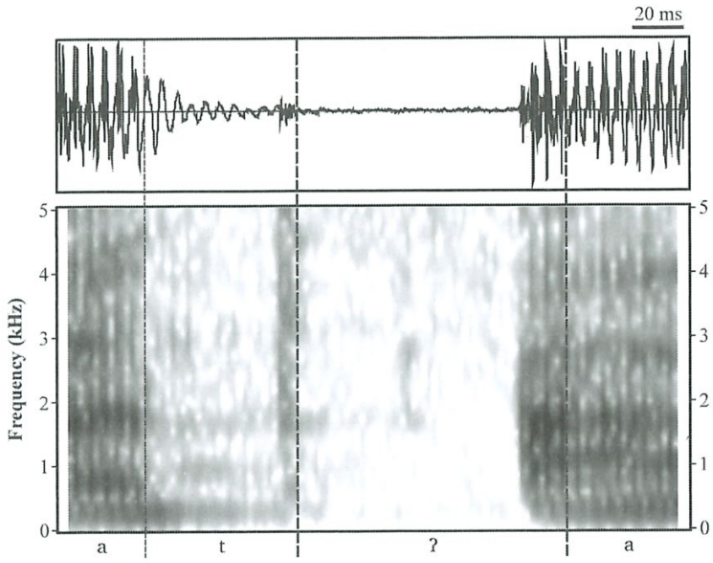

Figure 11.1: Sequence [at?a] showing the high-intensity periods belongs to release phase of the glottal stop

- Left boundary: Preglottalisation – as part of the glottal stop

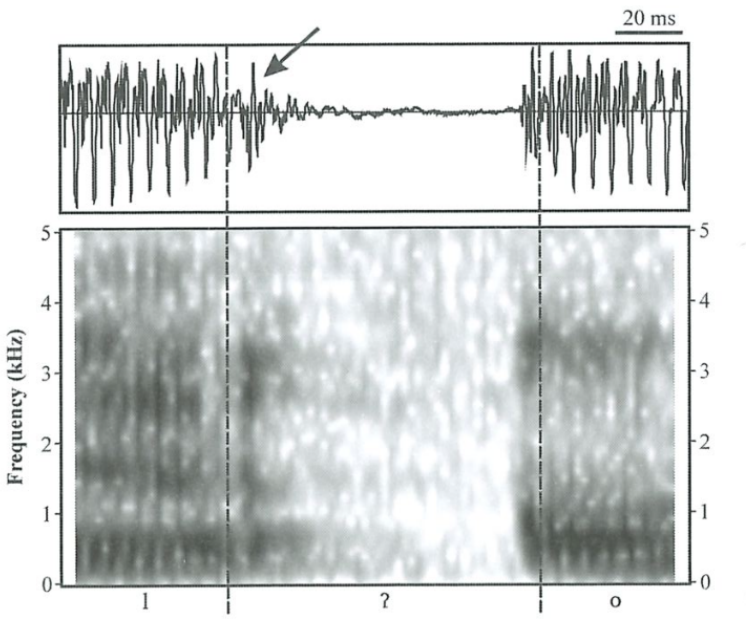

Figure 11.2: Sequence [lʔo] showing the last period before the closure with glottalisation

11.2 Creaky glottal stops

Creaky phonation is characterised by aperiodicity. Generally, any aperiodicity is regardde as part of the glottal stop (as shown in Figure 11.3), while it might result in short duration of the preceding vowel (as shown in Figure 11.4).

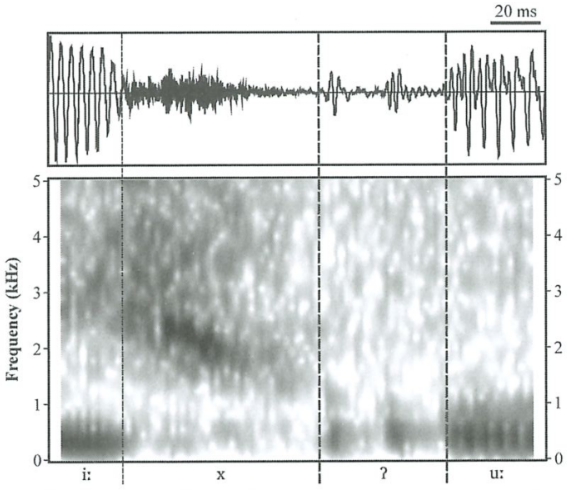

Figure 11.3: Sequence [i:xʔu:] showing two aperiodic pulses in the glottal stop

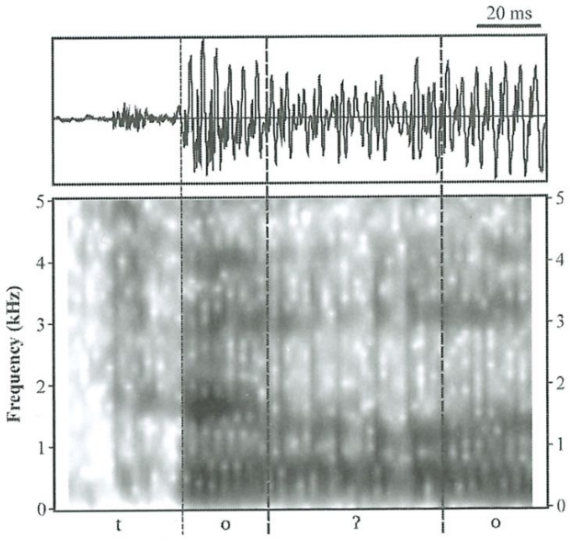

Figure 11.4: Sequence [toʔo] showing a very short [o] due to the extensive aperiodic component of the following glottal stop

12 Utterance beginnings and ends

12.1 Initial (Post-pausal) utterances

- Voiceless plosives/affricates – Stipulate a closure phase of 40-70 ms
- Initial devoicing of voiced plosives/sonorants – Stipulate a closure phase of 40-70 ms

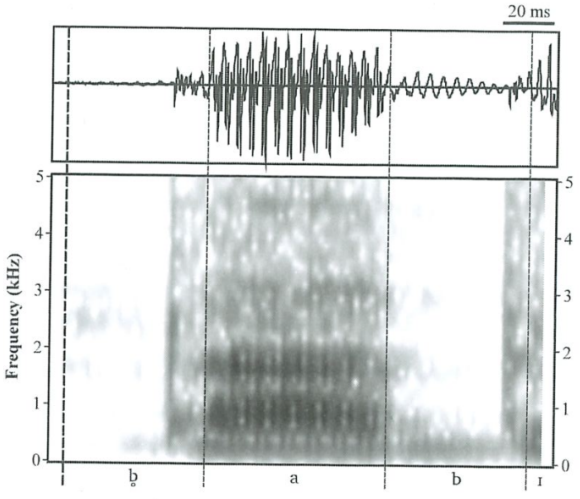

Figure 12.1: Sequence [b̥abɪ] with a devoiced [b̥], marking a hold phase at 40 ms

- Hard/Soft glottal onsets – apply rules of segmenting glottal stops

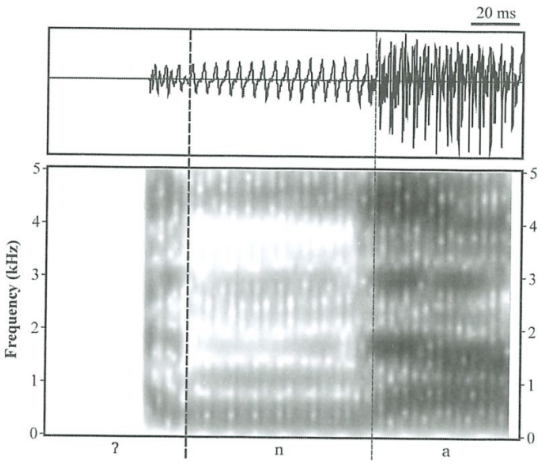

Figure 12.2: Sequence [ʔna] with with a hard glottal stop before the nasal

12.2 Final (Pre-pausal) utterances

We usually stick to The Full Formant Structure Rule to mark the end of the utterance. There are three cases regarding the utterance endings.

- Continuation of modal voicing
- Creaky or breathy voicing
- Epenthetic schwa-like element