Pavel Machač
& Radek Skarnitzl

Principles of Phonetic SEGMENTATION

Keywords:
- spectrogram
- formant structure
- acoustic cues
- waveform
- transition phase
- formant column
- speechsound combinations
- inter-labeller consistency
- boundary location
- spectral intensity
- careful listening

EDITION ERUDICA

EPOCHA PUBLISHING HOUSE

EDITION ERUDICA

EPOCHA PUBLISHING HOUSE

PRINCIPLES OF PHONETIC SEGMENTATION

NAKLADATELSTVÍ
EPOCHA

Pavel Machač & Radek Skarnitzl

PRINCIPLES OF PHONETIC SEGMENTATION

EPOCHA PUBLISHING HOUSE

Acknowledgements

We would like to express our gratitude to our colleague and friend, Jan Volín, who gave us the impetus to write this book, as well as valuable advice and experience. We would also like to thank Jana Heranová and Lucie Ondrušková for their insightful comments on an earlier version of the manuscript.

This book was written with the support of the grants MRTN-CT-2006-035561 (European grant, Sound to Sense), VZ MSM 0021620825 of the Czech Ministry of Education, and GACR 102/09/0989 (Chapters 10 and 12).

ISBN 978-80-7425-032-3

Contents

Motto: Everything has boundaries,
though often unclear.

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.

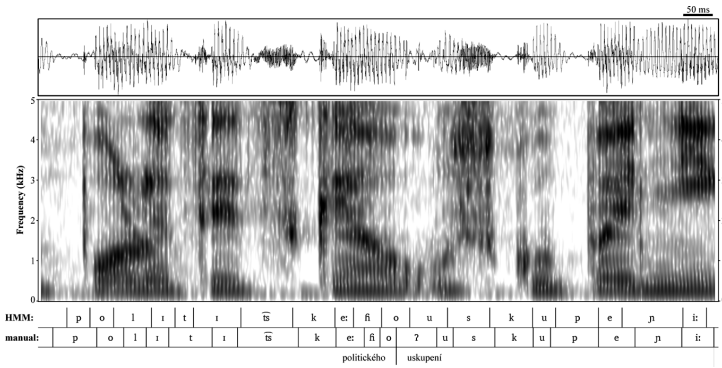

Figure 1.1. Comparison of HMM-generated and manually corrected segmentation of two Czech words.

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.