

# Phonetic Labeling and Segmentation of Mixed-Lingual Prosody Databases

Harald Romsdorfer and Beat Pfister

Speech Processing Group  
Computer Engineering and Networks Laboratory  
ETH Zurich

{romsdorfer,pfister}@tik.ee.ethz.ch

## Abstract

An automatic system for segmenting speech signals used for the training of statistical prosody models is presented. Starting from a canonical transcription, the system simultaneously delivers an accurate phonetic segmentation and the matched phonetic transcription indicating pronunciation variants.

Although the system is HMM-based, it uses only the speech signals of the prosody database which typically consists of a few hundred sentences with some 30 minutes total duration. Initial phone HMMs are generated with flat-start training using the canonical transcriptions of the sentences. Then iterative Viterbi search for best-matching pronunciation variants and HMM re-training is applied until convergence is attained.

## 1. Introduction

It is generally acknowledged that data-driven approaches to prosody control in text-to-speech (TTS) synthesis outperform rule-based approaches. In particular, artificial neural networks (ANNs) have been used successfully for fundamental frequency control and for generating phone durations.

Besides the topologies of these ANNs and their input and output coding, there are two prerequisites for this type of statistical approach to prosody control: First of all, the input to the ANNs must include all relevant linguistic information and secondly, a sufficiently large set of input/output data pairs representing all major prosodic phenomena has to be available for the training of the ANNs.

Our investigations have shown (e.g. in [1], [2]) that the required input to the ANNs can be derived from the so-called phonological representation which is shown for an example sentence in Figure 1. Although this phonological representation is rather slim, we could demonstrate that it is sufficient to generate synthetic melody and rhythm that listeners scored as very close to natural.

In order to satisfy the second of the above mentioned prerequisites, namely the availability of suitable training data, we need speech signals uttered by a speaker in a manner that we like the synthesizer to speak ultimately. The prosody of these speech signals must be described very accurately in terms of temporal structure (durations of phones and pauses), phonetic variation, syllable stress level, phrase boundaries and fundamental frequency.

Of course it is most desirable to have an automatic procedure to generate this description. Excluding the fundamental frequency estimation, such a procedure would comprise the following steps: (1) generating the phonological representation from the text; (2) specifying the pronunciation variation rules for the speaker at hand; (3) segmenting the speech signals.

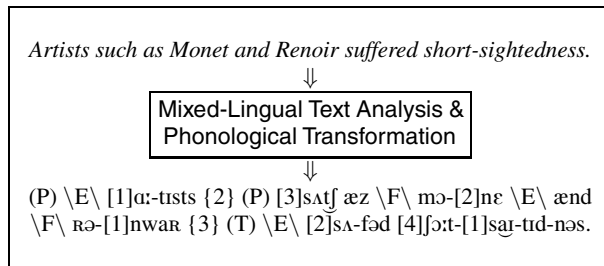


Figure 1: The phonological representation of a sentence, generated by the mixed-lingual text-analyzer (cf. [3] and [4]) includes phonetic symbols, language switching positions (marked between backslashes), syllable stress level (in square brackets), position and strength of phrase boundaries (in curly braces), and phrase types (in round brackets).

Item 1 has already been presented elsewhere (see [3] and [4]). This paper focuses on item 3 and presents a segmentation system which delivers the matched phonetic transcription and segmentation of the speech signals at the same time.

The structure of the paper is as follows: After the description of the segmentation problem in Section 2, the segmentation procedure is shown in Section 3. In Sections 4 and 5 the performed experiments and the achieved results are presented. Finally we draw some conclusions in Section 6.

## 2. The segmentation problem

Since manual phonetic segmentation of speech signals is tedious and tends to be neither consistent nor reproducible, automatic segmentation is widely used instead. There exist essentially two types of segmentation methods that use either dynamic time warping (DTW) or hidden markov models (HMMs). In DTW-based approaches synthetic speech signals with known phone boundaries are temporally aligned to the natural speech, whereas in HMM-based systems forced alignment is used.

Although the findings of other researchers do not clearly indicate the superiority of one or the other approach (see e.g. [5] and [6]), we think that the HMM-based one is more appropriate for our needs, as we will outline in the sequel.

We have experienced that for the training of prosody ANNs a speech database with only few hundred sentences per language is sufficient, provided that the speaking style is constant over all sentences and the accuracy of the annotations and of the segmentation is high enough. Our intention to attain a sufficiently high level of accuracy has resulted in an approach which takes the following considerations into account:

- We hypothesize that forced alignment with more specific phone HMMs (i.e. variance of Gaussians as small as possible) results in more precise phone segment boundaries. Thus, we need HMMs that incorporate only the variability of the data to be segmented. In other words: We should use the speech data to be segmented as training set, because we do not want to apply the HMMs to other signals. Since the training data is fairly limited, it is preferable to use context-independent phone models with few mixtures only.
- Speech signals used for the training of prosody models are generally spoken in a fluent or even lively manner and therefore deviate more or less from the standard pronunciation. A good segmentation has to account for such pronunciation variation. Most such variations are fairly regular, even across languages, and can appropriately be described by a small set of rules. For some speakers, however, additional rules may be necessary. In order to account for the fact that pronunciation variations happens also quite frequently across word boundaries (e.g. fusion of two identical phones across a word boundary), the rules are used to generate variants of the canonical phonetic transcriptions of the sentences (not of the entries in the pronunciation lexicon as e.g. in [7]). The best matching transcription and the corresponding phone boundaries can then be determined by Viterbi segmentation.
- In order to create more specific phone HMMs, i.e. to reduce the degradation from pronunciation variation, the adapted transcriptions can be used iteratively to train new HMMs and to generate new transcriptions of the sentences until the best solution is attained.
- Last but not least, we also want to segment mixed-lingual sentences with words (and phonemes) of several languages (see example in Figures 1 and 2).

The segmentation procedure described in the next section takes all these considerations into account.

### 3. Automatic segmentation procedure

The segmentation procedure consists of two stages: First, context-independent phone HMMs are trained using the canonical phonetic transcriptions of the sentences. Apart from optional pauses between words no pronunciation variants are considered in this stage.

For the second stage the pronunciation variation rules are applied to the canonical sentence transcriptions and a recognition network is generated for each sentence. It has to be emphasized that the network includes all pronunciations allowed by the rules.

Then a Viterbi search determines the most likely path through the networks and thus delivers an adapted phonetic transcription of each sentence. These new transcriptions are used to retrain the HMMs that are in turn used in the next iteration for the Viterbi search. The procedure stops when the adapted phonetic transcriptions do not change any more.

#### 3.1. Parameters estimation of HMM phone models

To account for the limited training data, we use context-independent three-state left-to-right phone HMMs with four Gaussian mixtures per state. Experiments have shown equal results for eight and inferior results for ten or more mixtures.

The features used are 12 mel-frequency cepstral coefficients, 12 delta coefficients, log energy and delta log energy.

These features are calculated over an analysis window of 15 ms length (for a female speaker) with a window shift of 2.5 ms. The temporal resolution of the segmentation is therefore limited to 2.5 ms which proved to be enough for our segmentation purposes.

The initialization of the phone HMMs is done using the so-called “flat start” training where initially all models are equal. Thus, no manually labeled training data is necessary to initialize the HMMs. Through repeated re-estimation using embedded Baum-Welch training with the given transcription (i.e. the canonical one in the first stage and successively adapted transcriptions in the second stage) we get optimized, speaker-dependent phone HMMs.

#### 3.2. Pronunciation variation rules

As mentioned above, the rules describing the pronunciation variations are either general, language-specific or speaker-specific (this distinction is not always clear, but anyway not relevant for the segmentation procedure). The rules used for the segmentation experiment described in Section 4 specify variations such as:

##### General rules:

- geminate reduction across word boundaries, but not across phrase boundaries:  
“flammable liquids” [flæməbl ɪkwɪdz] → [flæməblɪkwɪdz]
- deletion of plosive before another plosive, also across word boundaries:  
“conduct” [kɒndʌkt] → [kɒndʌt],  
“that person” [ðæt pɜːsn] → [ðæpɜːsn]
- vowel reduction in unstressed syllables:  
“sightedness” [saɪtɪdnɪs] → [saɪtɪdnəs]
- insertion of preplosive pause before voiced plosive (the preplosive pause is denoted by [>]):  
“sightedness” [saɪtɪdnɪs] → [sʌɪtɪ>dnɪs]
- aspiration of unvoiced plosives before vowels:  
“country” [kʌntri] → [kʰʌntri]

##### Language-specific rules:

- schwa deletion between plosives and nasals:  
“controlling” [kɒntrəʊlɪŋ] → [kɒntrəʊlɪŋ]
- schwa deletion before [r]:  
“covering” [kʌvəɪŋ] → [kʌvɪŋ]
- deletion of word final plosive after nasal:  
“student safety” [stjuːdnt seɪfti] → [stjuːdnseɪfti]
- deletion of [g] after [ŋ]:  
“English” [ɪŋɡlɪʃ] → [ɪŋlɪʃ]

##### Speaker-specific rules:

- deletion of word final [R] in French words:  
“Renoir” [Rɛnwɑː] → [Rɛnwa]
- insertion of homorganic fricative after unvoiced plosive:  
“controlling” [kɒntrəʊlɪŋ] → [kɒntʃrəʊlɪŋ]

The formalism of the pronunciation variation rules is illustrated by some examples in Table 1. Note that many of the above listed variations need more than one rule to be described.

Since most assimilation phenomena occur also or even primarily across word boundaries (if they are not coincident with strong phrase boundaries), the rules are applied to the phonological representation of the complete sentences (as shown in Figure 1). The result is a recognition network per sentence.

|    |                    |    |                    |
|----|--------------------|----|--------------------|
| a) | k / k <sup>h</sup> | => | _ %Vowel ;         |
| b) | NULL / >           | => | _ %VoicedPlosive ; |
| c) | %Plosive / NULL    | => | _ [ # ] %Plosive ; |

Table 1: Examples of rewrite rules defining some of the pronunciation variants listed above. The left side specifies the rewrite operation, the right side the phonetic context triggering this rule using ‘\_’ to denote the rewrite position. ‘NULL’ specifies the empty symbol and ‘#’ the word boundary. ‘%’ indicates a set identifier. In rule a) [k] before vowels is optionally aspirated. Rule b) specifies optional insertion of a preplosive pause (>) before voiced plosives. Rule c) denotes deletion of a plosive before another plosive either within a word or across word boundaries (‘#’). As phrase boundaries are denoted using a different symbol this rule is not triggered at phrase boundaries.

It is worth to mention that all types of pronunciation variation rules are equally and simultaneously applied and in particular no weighting of the rules, as e.g. in [8], is needed here.

For the description of all pronunciation variations listed above we needed a set of only 74 rewrite rules, some of them covering French inclusions. This rule set was applied in the experiments of Section 4.

## 4. Experiments

We used a single speaker corpus of 33 minutes length consisting of 402 English sentences containing some French inclusions. The corpus was recorded to train UK English prosody models for our polyglot TTS synthesis system. The sentence lengths range from 1.3 seconds up to 20 seconds. On average a sentence contains about 24 syllables ranging from 3 to 81 syllables.

The graphemic sentences were transcribed using the mixed-lingual transcription module of our polyglot TTS synthesis system generating the phonological representations (as shown in Figure 1). From these phonological representations the recognition networks for the second stage of the segmentation system were generated using the pronunciation variation rules described above.

After each iteration of the second stage (see Section 3) the insertions, deletions and replacements of segments relative to the preceding iteration were counted. If the procedure converges (this was not clear at the beginning) these figures are expected to decrease more or less steadily. The procedure stops as soon as the number of changes stays below a certain threshold. For the English UK corpus the segmentation stopped after 14 iterations as shown in Table 2.

## 5. Results

The performance of the automatic segmentation system was assessed in terms of segment boundary accuracy and compared with the results of a manual segmentation. For that purpose a subset of 30 sentences containing about 1800 segments was manually segmented and labeled by a professional phonetician.

It goes without saying that the phonetician and the automatic system did not use the same criteria for placing the segment boundaries. As can be seen from the example in Figure 2, large differences occur e.g. at the boundaries between the plosive and the fricative of affricates such as [ts] and [tʃ]. The largest deviation between manual and automatic segmentation resulted at the boundary between a pause and the preceding phone.

| Iteration | Insertions | Deletions | Replacem. | Total |
|-----------|------------|-----------|-----------|-------|
| 1         | 560        | 1467      | 1216      | 3250  |
| 2         | 234        | 261       | 291       | 786   |
| 3         | 111        | 77        | 117       | 308   |
| 4         | 65         | 48        | 77        | 192   |
| 5         | 49         | 15        | 47        | 111   |
| ⋮         | ⋮          | ⋮         | ⋮         | ⋮     |
| 12        | 4          | 3         | 5         | 12    |
| 13        | 5          | 0         | 0         | 5     |
| 14        | 3          | 2         | 0         | 5     |

Table 2: Number of insertions, deletions and replacements found after each iteration (segmentation and HMM retraining). The corpus contains about 31000 segments. In the 14th iteration the number of variations stayed below the given threshold indicating that the procedure converged.

Even though the criteria for placing boundaries are different we have assessed the results of the automatic system against the manual segmentation, because we want to replace the latter by the former. But the different criteria explain why our segmentation is less accurate than e.g. the one reported in [8]. Nevertheless, the segmentation accuracy as given in Table 3 fulfills our requirements arising from the training of prosody models.

| Deviation | Initial segm. | Iteration 1 | Iteration 14 |
|-----------|---------------|-------------|--------------|
| < 5 ms    | 38.0 %        | 44.7 %      | 44.3 %       |
| < 10 ms   | 65.5 %        | 68.4 %      | 68.1 %       |
| < 15 ms   | 79.2 %        | 80.7 %      | 81.1 %       |
| < 20 ms   | 84.9 %        | 86.9 %      | 86.9 %       |
| < 25 ms   | 89.2 %        | 91.9 %      | 91.2 %       |
| < 30 ms   | 91.6 %        | 93.8 %      | 93.7 %       |
| < 40 ms   | 95.0 %        | 96.2 %      | 96.0 %       |
| < 60 ms   | 97.5 %        | 97.9 %      | 97.5 %       |
| < 200 ms  | 99.7 %        | 99.8 %      | 100 %        |

Table 3: Percentage of automatically set boundaries that deviate less than 5 ms, 10 ms, etc. from the manually set boundaries after the initial segmentation (first stage of the procedure) and after the first and the 14th iteration of the second stage. In this table, only the boundaries between equal phones in the manual and automatic segmentation have been considered.

Another quality measure for the system is to compare the number of insertions, deletions and replacements between the manual and the automatic phonetic segmentation and labeling. As shown in Table 4, 283 or about 16 % of the approximately 1800 segments of the canonical transcription were modified in the manual labeling. The automatic procedure detected 165 or 58.3 % of these variations.

|               | Insertions | Deletions | Replacem. | Total |
|---------------|------------|-----------|-----------|-------|
| Manual        | 83         | 111       | 89        | 283   |
| Detected      | 50         | 43        | 72        | 165   |
| Not detected  | 33         | 68        | 17        | 118   |
| Add. Inserted | 14         | 38        | 21        | 73    |

Table 4: Number of pronunciation variants of the manual labeling in terms of insertions, deletions, replacements and in total. Also, the number of variations that were detected, not detected or additionally inserted by the automatic procedure are denoted.

## 6. Discussion and conclusions

The presented HMM-based segmentation system is able to do an accurate phonetic segmentation, even if the available speech data is fairly limited. A collection of sentences used for the training of TTS prosody models represents such a case. The segmentation system is well adapted to the opportunities and requirements of this case:

- Before the segmentation can start, the phonological representation of the sentences (text) has to be generated. Therefore, the main prerequisite is the availability of a text analyzer for the languages given by the set of sentences. Since such an analyzer is used for the TTS system anyway, it is assumed to be given (i.e. we first realize the text analyzer and then the prosody models).
- The phone HMMs used for forced alignment segmentation are trained with the set of few hundred sentences only. In particular, there is no other speech data necessary (e.g. for the training of a phonetic recognizer).
- Pronunciation variation is specified by a set of rules. The number of these rules is typically fairly small, even though inter-word phenomena are also included (which are quite frequent and not considered in most other systems). The majority of the rules describe rather general coarticulation phenomena and rules for language- or speaker-specific phenomena can easily be added.
- In contrast to segmentation used for unit selection TTS, accuracy of phone boundaries is generally more critical in our case (diphone concatenation-based TTS synthesis) than the detection of all pronunciation variants. E.g. frequent assimilations such as [nb] → [mb] or [nk] → [ŋk] do not matter at all.
- Last but not least, the system is practical, particular when targeting TTS for a diversity of languages: A few hundred sentences for each language are required only and the system even can process mixed-lingual sentences.

In conclusion, the presented automatic segmentation system works accurately enough to replace the earlier used manual procedure.

## 7. Acknowledgments

We cordially thank Alexis Wilpert for the carefully done manual segmentations used for assessing the automatic segmentation.

This work was partly supported by the Swiss National Science Foundation in the framework of NCCR IM2.

## 8. References

- [1] C. Traber. *SVOX: The Implementation of a Text-to-Speech System for German*. PhD thesis, No. 11064, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 7, ISBN 3 7281 2239 4), March 1995.
- [2] M. Riedi. *Controlling Segmental Duration in Speech Synthesis Systems*. PhD thesis, No. 12487, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 26, ISBN 3-906469-05-0), February 1998.
- [3] B. Pfister and H. Romsdorfer. Mixed-lingual text analysis for polyglot TTS synthesis. In *Proceedings of Eurospeech'03*, pages 2037–2040, Geneva, September 2003.
- [4] H. Romsdorfer and B. Pfister. Multi-context rules for phonological processing in polyglot TTS synthesis. In *Proceedings of Interspeech 2004 – ICSLP*, pages 737–740, Jeju Island (Korea), October 2004.
- [5] F. Malfère, D. Deroo, T. Dutoit, and C. Ris. Phonetic alignment: Speech synthesis-based vs. Viterbi-based. *Speech Communication Nr. 40*, 40(4), June 2003.
- [6] R. Kominek, C. Bennett, and A. Black. Evaluating and correcting phoneme segmentation for unit selection synthesis. In *Proceedings of Eurospeech'03*, Geneva, September 2003.
- [7] Y.-J. Kim, A. Syrdal, and A. Conkie. Pronunciation lexicon adaptation for TTS voice building. In *Proceedings of Interspeech 2004 – ICSLP*, Jeju Island (Korea), October 2004.
- [8] J. van Santen and R. Sproat. High-accuracy automatic segmentation. In *Proceedings of Eurospeech'99*, Budapest (Hungary), 1999.

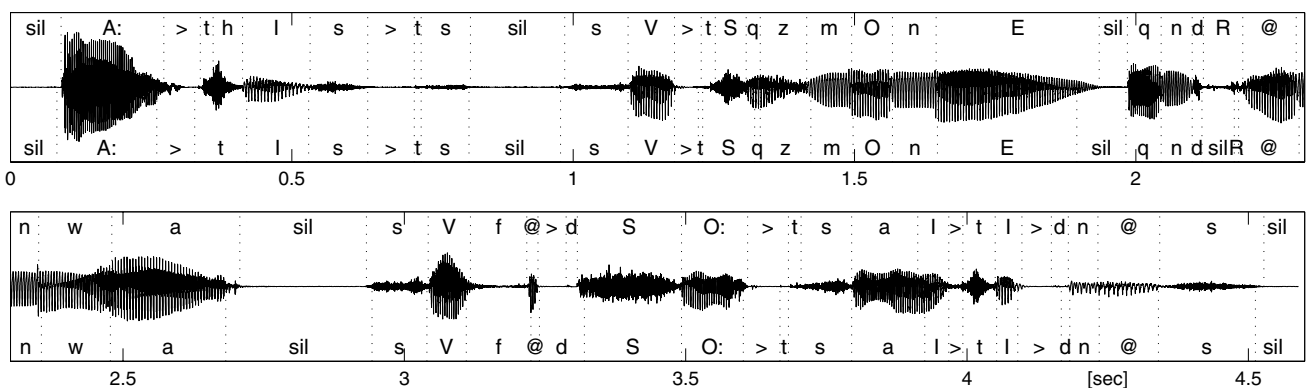


Figure 2: Comparison of manual and automatic segmentation of a speech signal with the wording given in Figure 1. The upper and lower labels correspond to the manual and automatic segmentation, resp. The phonetic labels largely follow the SAMPA definitions.