# Boundary Deviation Of Phonemes In Automatic Segmentation Systems - A Cross Language Study

**Nicole Beringer**

Institut fuer Phonetik und sprachliche Kommunikation
University of Munich
Schellingstr. 3
80799 Munich
Germany
beringer@phonetik.uni-muenchen.de

## ABSTRACT

This work aims to analyze the boundary deviation of phonemes in automatic segmentation systems of German, American English and Japanese compared to manually set boundaries and across language-comparable phonemes. It further gives the deviation per language-specific phoneme of the automatic segmentation compared to the reference segmentation for matching transcripts and a detailed intra-language analysis and cross-language analysis. Given these results an approach is made to improve segment boundaries in automatic segmentation systems.

## 1 Introduction

Based on previous work ([1], [2], [3], [4], [[5]) where it was shown that automatic segmentation systems can reach a correspondence of phonemes of up to 97.5% compared to manually segmented data (dependent on a specific domain (Munich Automatic Segmentation system - MAUS([1]) or domain-independently ([3], MAUSER [4])) we observed, however, that the segmentation boundaries of phonemes vary due to the acoustic models independently of the used HMM algorithm.

The aim of the paper is to give the phonemes most prominent in deviation for German, American English and Japanese, compare them cross-linguistically and find an explanation for it.

The following section gives a short description of the acoustic training and the phonemic characteristics of phonemes comparable across languages. Section 3 analyzes this shift of phoneme boundaries based on German, English and Japanese Verbmobil segmentations which seems to be caused due to an inherent processing problem within Hidden Markov Modelling. Section 4 gives a detailed explanation of why several phonemes have more prominent deviations than others (intra-language analysis) and why comparable phonemes of the three languages do not always show a comparable deviation (cross-language analysis). An approach to improve segment boundaries in automatic segmentation systems is made in the last section of the paper.

## 2 Basic Specifications

### 2.1 Database

The investigation presented in this paper uses a database of unprompted speech - the VERBMOBIL corpus ([7]). The database-scenario deals with scheduling appointments which are real-life-situations with contemporary speech. It consists of three language portions: German, American English and Japanese.

The German VERBMOBIL portion contains sufficient speech data for training and testing (35136 turns[1]). The English portion contains 20547 turns and the Japanese 34302 turns.

Due to task instructions speakers had to ask for departures or arrange an appointment with a business partner. The “formal situation” - setup makes sure that speech contains fewer and weaker regional variants than it would contain if personal affairs were discussed because people tend to hyper-articulate in these situations.

### 2.2 Acoustic Models

The acoustic modelling within the MAUSER-system is based on the Hidden-Markov-Toolkit. For German 474 manually segmented turns were used for training [6]. The number of English turns used for training was 14875, for Japanese 23536 turns. For both languages we used a flat-start in the training process.

[1] One turn in the VERBMOBIL database has about 22.8 words in average

ISBN 1-876346-48-5 © 2003 UAB

All models are speaker-independent. The models used in this experiment were trained to more than 700 speakers over all three languages. Therefore, rather broad statistical distributions in the HMM states that might overlap an inter-lingual variability are expected.

### 2.3 Phonemic Characteristics

None of the phonemes in consideration are split by means of diacritics in the HMM-training. Therefore, phenomena like aspiration, glottalization, fricativization, voicing, vocalization, monophthongization and centralization are not considered in the acoustic models, and the model for a cross-language-equivalent phone[2] (table 1) includes these phenomena on a bigger or smaller scale.

Comparing the standard SAM-PA phone systems (about 45 phonemes per examined language) correlated to the used HMMs it can be seen that about half of the equivalent phones of all languages are the same. Bilingual comparisons can have at most two thirds of equivalent forms as can be observed with German-English.

| Languages | equivalent phones | number |
|---|---|---|
| ger-eng-jap | @, N, S, a, b, d, e, g, h, | 18 |
| | j, k, l, m, n, p, s, t, z | |
| ger-jap | @, C, E, N, S, a, b, d, e, g, | 20 |
| | h, j, k, l, m, n, p, s, t, z | |
| ger-eng | @, I, N, S, U, a, aI, aU, b, | 28 |
| | d, e, f, g, h, i:, | |
| | j, k, l, m, n, p, p:, r | |
| eng-jap | @, N, S, a, b, d, e, g, h, | 18 |
| | j, k, l, m, n, p, s, t, z | |

**Table 1:** Comparison of equivalent phones in the language specific SAM-PA. ger stands for German, eng for English, jap for Japanese.

Considering the deviations of phoneme boundaries it can be stated that especially phonemes with variing aspects of aspiration or variing length in the acoustic models deviate significantly as will be discussed in the following section.

## 3 Boundary Shifts

Table 2 gives the deviation per language-specific phoneme of the automatic segmentation compared to the manual references for matching transcripts.

As can be seen, German vowel boundaries mostly show a deviation of under 15m,s except /Y/ and /i:/ with a deviation of more than 25ms, and /a/, /2:/, /U/ and /aI/ with a deviation under 5ms. In contrast, the segment boundaries of voiceless plosives deviate compared to the manually set boundaries by up to 55ms. Most fricatives and semivowels have a deviation of about 25ms. The rest of the German consonant boundaries lie between 5 and 10ms of deviation.

American English vowels and diphthongs deviate mostly by about 5 to 10 ms. Exceptions are /V/ (just over 10ms), /aI/ (little above 15ms), /@/ and /aU/ (both above 30ms). Unlike the German plosives English plosives showed a deviation of under 20ms in the experiments except /g/, which deviates by 25ms. Affricates and fricatives also have lower deviations than the comparable German ones (often under 10ms). Exceptions are /T/ (over 30ms), /D/ (almost 20ms), /h/ (left boundary 30ms). Semivowels, liquids and nasal consonants have a deviation under 10ms.

Despite good training values of the acoustic models, Japanese phonemes in our experiments show on average higher deviations of segment boundaries than the other two languages (usually between 15ms and 25ms), although there is no higher deviation than 25ms. Exceptions here are /p/,/b/,/dz/,/h/,/l/ and /o/ with deviations of segment boundaries between 15ms and 25ms.

Considering the English and Japanese word ends (#) a deviation of about 10ms (English) to 15ms (Japanese) was found. The right boundaries of pauses deviate almost 15%. The detection of the left boundary was better for both languages: around 2% (English) and 12% (Japanese).

Given these results it can be assumed that deviation is caused due to an inherent processing problem within Hidden Markov Modelling usually for those phonemes which show aspiration or fricativization. The fact that corresponding phonemes of the investigated languages also vary across languages as well as observations with artifical networks, where we cannot observe a similar deviation, strengthen this assumption.

## 4 Analysis Of Deviation

### 4.1 Intra-language Analysis

Table 2 enumerates all phonemes with their mean left and right boundary shift.

**German**

Considering the consonants first, what strikes most is that all left voiceless plosive boundaries are set earlier automatically than manually (/p/ -27.73%, /t/ -30.8%, /k/ -10.94%). Except for the bilabial voiceless plosive, all right boundaries are also strongly shifted to the left (-54.41 /t/ and -30.19 /k/). An explanation could be the effect aspiration has on voiceless plosives in word boundary position, their most prominent positions in

[2] Cross-language-equivalent phone: a phone which is represented by the same IPA symbol in all languages in question.

ISBN 1-876346-48-5 © 2003 UAB

| | | *German* | | *English* | | *Japanese* | |
|---|---|---|---|---|---|---|---|
| phoneme class | Label | mean-begin | mean-end | mean-begin | mean-end | mean-begin | mean-end |
| diphthongs | aI | 1.07 | 0.54 | 15,92 | 14,09 | | |
| | eI | | | 0,97 | 4,03 | | |
| | aU | -9.08 | 10.27 | 0,34 | 0,77 | | |
| | e@ | | | 1,02 | 0,95 | | |
| | @U | | | 26,67 | 30,63 | | |
| | OY | 14.75 | 10.42 | | | | |
| dark vowels | A | | | | | 16,24 | 16,35 |
| | U | 3.73 | -1.28 | 1,41 | 1,49 | | |
| | u: | 8.5 | 9.03 | 2,42 | 0,17 | | |
| | a | -0.62 | -0.36 | | | | |
| | 6 | -8.75 | -19.8 | | | | |
| | V | | | 8,61 | 11,67 | | |
| central vowels | e | | | -2,19 | -1,36 | 17,24 | 16,97 |
| | E | 9.04 | 6.56 | | | | |
| | e: | 2.81 | -16.92 | | | | |
| | @ | -10.56 | -18.04 | 29,21 | 31,77 | | |
| | 2 | 2.49 | -2.83 | | | | |
| | 9 | 12.29 | 9.54 | | | | |
| | O | | | 6.23 | 8.15 | | |
| | O: | 0,62 | 4,62 | | | | |
| | o | | | | | 12,77 | 12,33 |
| | o: | 5.44 | 1.5 | | | | |
| | 3: | | | 0,22 | 1,26 | | |
| | { | | | -8,3 | -4,25 | | |
| high vowels | i | | | | | 16,45 | 15,72 |
| | 1 | | | | | 18,04 | 17,86 |
| | W | | | | | 17,29 | 16,46 |
| | I | 12.53 | 6.48 | 5,15 | 6,26 | | |
| | i: | 5.54 | -27.19 | 3,73 | 4,49 | | |
| | Y | 27.69 | 28.21 | | | | |
| semivowels | j | 31.31 | 4.14 | 4,15 | 5,26 | 19,65 | 19,3 |
| | w | | | 6,81 | 0,03 | | |
| liquids | R | | | | | 16,58 | 16,45 |
| | r | -2.82 | -2.78 | -0,98 | -3,27 | | |
| | l | 7.46 | -4.17 | 6,72 | 0,3 | 0,88 | 1,33 |
| nasals | m | 7.43 | 9.16 | 8,55 | 9,04 | 16,46 | 16,48 |
| | n | -2.87 | -2.4 | 3,71 | 2,75 | 20,84 | 20,91 |
| | N | 3.36 | 1.38 | -6,92 | -9,15 | 15,8 | 15,49 |
| plosives | ? | 15.3 | 7.18 | | | | |
| | p | -27.73 | -8.72 | 18,45 | 8,89 | 15,52 | 10,7 |
| | b | -7.57 | 5.51 | 6,37 | 0,94 | 9,08 | 8,61 |
| | t | -30.8 | -54.41 | 10,83 | 12,27 | 23,65 | 20,64 |
| | d | 6.59 | 2.82 | -6,71 | -7,04 | 22,13 | 22,51 |
| | k | -10.94 | -30.19 | 19,33 | 9,95 | 20,21 | 19,74 |
| | g | -0.89 | 3.76 | 26,79 | 23,21 | 22,36 | 23,23 |
| affrikates | *ts* | | | | | 16,69 | 17,02 |
| | *dz* | | | | | -11,87 | -13,94 |
| | tS | | | 12,96 | 12,41 | | |
| | dZ | | | -2 | -2,13 | | |
| frikatives | T | | | 31,42 | 35,26 | | |
| | D | | | 19,5 | 20,08 | | |
| | s | 3.31 | 1.79 | 2,02 | -0,01 | 16,31 | 16,23 |
| | z | 14.51 | 1.69 | -3,53 | -5,3 | 24,56 | 24,45 |
| | S | 5.92 | 3.6 | 1,77 | -1,45 | | |
| | f | 2.08 | 9.03 | 17,08 | 5,9 | | |
| | v | 15.5 | 25.16 | 0,5 | -4,38 | | |
| | h | 3.82 | 1.48 | 26,79 | 13,52 | 12,03 | 11,29 |
| | C | 12.21 | 3.82 | | | 23,14 | 23,1 |
| | x | 22.63 | 3.35 | | | | |

**Table 2:** Mean deviation of segment boundaries (automatic segmentation vs. manual segmentation) in ms

ISBN 1-876346-48-5 © 2003 UAB

German: the silence after the burst is recognized as a pause between words, the plosive in question is shifted.

Voiced fricatives and /C/ and /x/ show a big shift to the right for the left boundary. Only /v/ shifts the right boundary more than 20ms on average. Except for /C/ and /x/, which do not show clear features for boundary detection this phenomenon can be manipulated by the surrounding vowels which tend to evaporate over voiced neighbours.

The semivowel /j/ shows a big shift to the right for the left boundary. The right boundary is shifted about 5ms on average. The neighbouring vowels shift the boundaries here as well.

High vowels show remarkably shifts to the right only for the left boundaries of /I/ and /Y/ and the right boundary of /Y/.

The left boundary of the central vowel /9/ is set later in the signal by the automatical segmentation, whereas /@/ is shifted to the left with both boundaries.

Only the right boundary of /6/ is extremely shifted to the left.

The diphthongs /OY/ and /aU/ show right shifts for both segment boundaries (former) and the right segment boundary (latter) respectively.

The shift of vowel boundaries is mostly found with surrounding voiced consonants. Why this can be found only with non-back vowels and especially with lax ones and schwa in German remains to be examined.

**English**
All remarkable English boundary shifts are to the right which are for the diphthongs /aI/ and /@U/, /@/, the affricate /tS/, the fricatives /T/, /D/ and /h/ and the voiceless plosives /t/ and /k/ for both segment boundaries and the plosive /p/ as well as the fricative /f/ only the left boundary. Plosives, fricatives and affricates are characterized by aspiration or friction which influences the right boundaries. The beginning of the phoneme is often hard to detect and can be mixed up with neighbouring phonemes.

**Japanese**
For Japanese only two phonemes show deviation under 10ms namely /l/ (0.88ms left, 1.33ms right) and /b/ (9.08ms left, 8.61ms right). As has been stated for German, the HMMs of vowels tend to extend in their neighbouring phones. Plosives, mostly fricativized or at least aspirated, affricates and fricatives vary in their boundary settings.

For the Japanese liquid the deviation is remarkably low which is a point to discuss.

### 4.2 Cross-language Analysis

Comparing the three languages it can be observed that all three languages show high deviations with the right segment boundaries of voiceless plosives, high deviations with the left segment of voiceless plosives, high deviations with fricatives and affricates and deviations with /@/ (English, German) and /j/ (German, Japanese). Also affricates generally show fairly high boundary deviations.

As for the vowels, HMMs seem to be extended into the neighbouring segments. This can be found especially with voiced fricatives and semivowels. For the voiceless plosives the number of word boundary plosives is higher than in-word plosives which means that they are statistically overweighted. When combined with aspiration or friction this could lead to shifting the boundaries to the left (German) or to the right (English, Japanese).

## 5 Future Work

One refinement approach we currently implement in our automatic segmentation system is to minimize the shift by re-shifting the boundaries with the normalized mean deviation we found for each phoneme class.

Apart from that, we search for cross-linguistically comparable HMM states and transitions to improve the acoustic modelling.

## REFERENCES

[1] Kipp, Wesenick, Schiel: Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech; EUROSPEECH 1997

[2] Wesenick : Automatic generation of German pronunciation variants. ICSLP 1996.

[3] Beringer, N.; Schiel, F. (1999) Independent Automatic Segmentation of Speech by Pronunciation Modeling. Proc. of the ICPhS 1999. San Francisco. August 1999. pp. 1653-1656

[4] N. Beringer (2002): Regeladaptive kategoriale Analyse von Spontansprache. (D). Doctoral thesis. University of Munich. Published by Shaker Verlag, Aachen, Germany, ISBN 3-8322-0267-6

[5] Andreas Kipp, Maria Barbara Wesenick, Florian Schiel (1996): Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora. in: Proceedings of the ICSLP 1996, Oct 1996, Philadelphia, pp. 106 109.

[6] K. Kohler. Labelled data bank of spoken standard German the Kiel corpus of read/spontaneous speech. In *Proceedings Of The ICSLP*, Volume 3, Philadelphia, 1996.

[7] K. Weilhammer, F. Schiel, U. Reichel. Multi-Tier Annotations in the Verbmobil Corpus. In *Proc. of the LREC 2002.*