Same same but different – An acoustical comparison of the automatic segmentation of high quality and mobile telephone speech

Christoph Draxler¹, Hanna S. Feiser¹,²

¹Institute of Phonetics and Speechprocessing, Ludwig-Maximilian University Munich, Germany
²Bavarian State Criminal Police Office, Munich, Germany

Abstract

In this paper we present a comparison of the performance of the automatic phonetic segmentation and labeling system MAUS [1] for two different signal qualities. For a forensic study on the similarity of voices within a family [2], eight speakers from four families were recorded simultaneously in both high bandwidth and mobile phone quality. The recordings were then automatically segmented and labeled using MAUS.

The results show marked effects on the segment counts and durations between the two signal qualities: for the mobile phone quality, the segment counts for fricatives were much lower than for high quality recordings, whereas the segment counts for plosives and vowels increased. The segment duration of fricatives was much lower for mobile phone recordings, slightly lower for the front vowels, but quite much longer for the back and low vowels.

Index Terms: forensic analysis, same-sex siblings, speech database, automatic segmentation, acoustical analysis

1. Introduction

The automatic processing of speech in the context of large speech databases has achieved significant progress during recent years. It is now possible, given an orthographic transcript of an utterance, to automatically align a text with the audio signal, or to generate a fine phonetic segmentation and labeling which even takes into account coarticulatory effects. However, in general this works reliably only for high quality signals; with noisy or compressed audio signals, the results deteriorate.

In the study presented here we systematically compare the performance of the MAUS system for read sentences for high bandwidth and mobile phone quality. These recordings were performed in the context of a forensic study on the similarity of voices in families – how similar are the voices of two brothers within a given family? With this setup, recording conditions were closely controlled, the only difference was the transmission channel. In forensic daily practice, this setup will quite likely show marked differences – for example in fricatives and vowels increased. The segment duration of fricatives was much lower for mobile phone recordings, slightly lower for the front vowels, but quite much longer for the back and low vowels.

1. Method

The speech database consists of 8 male speakers aged 20-31 years from four families. All speakers grew up in the area in and around Munich in Bavaria. Each speaker read 100 phonetically rich sentences from the Berlin Corpus and 20 minimal pairs in a movie fragment they each had seen prior to the dialog recording.

Following the setup of the DyVis [6] and the Pool [7] corpora, speakers were recorded in separate rooms using both high quality microphones and mobile phones. The high bandwidth recordings were made with a Neumann TLM 103 P48 dynamic microphone at 44.1 kHz sample rate and 16 bit quantization using the SpeechRecorder software [8]. At the same time, the speakers were recorded via mobile phone using Nokia 1680 and 2220 handsets respectively connected to an ISDN server. The signal quality of the mobile phone recordings thus is 8 kHz sample rate with 8 bit alaw quantization. For further processing, the quantization of the mobile phone recordings was converted to 16 bit linear PCM. Fig. 1 shows a sample segmentation of the word haben (to have) for both high quality and mobile phone recordings for the same utterance.

For the automatic segmentation and labeling, the web service version of the MAUS system was used [9]. The phoneme models were trained on high bandwidth speech with the German SAM-PA phoneme inventory. A standard right shift of the boundaries to the next 10ms is applied in a uniform way. For every recording, MAUS returned both the canonical form (i.e. citation pronunciation) of the words in the utterance, and a phonetic segmentation in Praat TextGrid file format. This segmentation takes into account coarticulatory effects, e.g. the @-elision in German syllables ending on -en such as /z a: g @ n/ vs. /z a: g n/.

The TextGrid files were then read into an SQL database
Figure 1: Sample signal for high quality a) and mobile b) recordings of the same utterance and the phoneme segments of haben.

Table 1: Segment count for phonemes

<table>
<thead>
<tr>
<th>code</th>
<th>quality</th>
<th>type</th>
<th>tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentences</td>
<td>mobile</td>
<td>44</td>
<td>16904</td>
</tr>
<tr>
<td></td>
<td>high quality</td>
<td>44</td>
<td>17003</td>
</tr>
<tr>
<td>minimal pairs</td>
<td>mobile</td>
<td>29</td>
<td>10351</td>
</tr>
<tr>
<td></td>
<td>high quality</td>
<td>29</td>
<td>10408</td>
</tr>
</tbody>
</table>

Table 2: Different automatic segmentations for the word haben.

<table>
<thead>
<tr>
<th>word</th>
<th>phoneme</th>
<th>mobile count</th>
<th>high quality count</th>
</tr>
</thead>
<tbody>
<tr>
<td>haben</td>
<td>h</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>a:</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>@</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>11</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 3: Counts for the phoneme classes by signal quality in the read sentences

<table>
<thead>
<tr>
<th>Phoneme class</th>
<th>mobile count</th>
<th>high quality count</th>
</tr>
</thead>
<tbody>
<tr>
<td>approximant</td>
<td>665</td>
<td>661</td>
</tr>
<tr>
<td>diphthong</td>
<td>693</td>
<td>683</td>
</tr>
<tr>
<td>nasal</td>
<td>2316</td>
<td>2364</td>
</tr>
<tr>
<td>fricative</td>
<td>3328</td>
<td>3684</td>
</tr>
<tr>
<td>plosive</td>
<td>3738</td>
<td>3667</td>
</tr>
<tr>
<td>vowel</td>
<td>5104</td>
<td>4888</td>
</tr>
</tbody>
</table>

system. The database contains a total of 54,666 phoneme segments, in 13488 orthographic word and canonical form segments.

Note that because MAUS assumes a hierarchical structure of elements in the different annotation tiers, i.e. a word has one canonic form and a canonic form may have many phonemes, this hierarchical structure is preserved in the segment table (technically, this is achieved by having a foreign key reference within the segment table).

For the statistical computations, the software R was used with the RDBMS interface library RPostgreSQL.

3. Analyses

3.1. Type and token counts

All speakers produced the same read utterances and hence the database contains the same orthographic word forms and canonical forms for every speaker and both recording qualities (341 word form or canonical form types and 4148 tokens for the sentences, 23 types and 2560 tokens for the minimal pairs).

However, for phoneme segments, there are differences: the inventory is the same, but the token counts are different. For both the sentences and the minimal pairs, there are more phoneme segments in the mobile phone recordings than in the high bandwidth recordings (Table 3.1).

In some words, phonemes are replaced by other phonemes due to coarticulation, e.g., /k/ by /h/ in gesagt (/g @ z a: k ə/; past tense of the verb to say). These replacements occur both in mobile phone and in high bandwidth recordings, but their counts differ. For example, for mobile phone speech, /k/ is used 396 times, and /h/ 244 times, but 446 and 192 times respectively for high quality speech in the word gesagt.

Other words have different segment counts for mobile and high bandwidth signal quality, e.g. the auxiliary verb haben (to have) which has only three phoneme distinct phoneme labels for high bandwidth quality, but six distinct phonemes for mobile phone quality – see Table 2.

From the 341 word forms, 221 (= 64.81%) have the same count of distinct phonemes for high quality and mobile phone speech, 103 (= 30.2%) have one different phoneme, 13 (=3.81%) have two, and 4 (= 1.17%) have three or more different phonemes.

Grouped by phoneme classes, it becomes clear that mainly the phoneme counts for fricatives, plosives, and vowels differ (Table 3).

3.2. Segment Durations

In the remainder of the paper, only the read sentences will be considered because they cover all German phonemes.
Due to the hierarchical annotations, the duration of the orthographic words and canonical forms is determined by the sum of the segment durations of the corresponding phoneme segments.

The total duration of the sentence segments is 1200.81s for the high quality recordings, and 1239.87s for the mobile phone recordings.

For high quality recordings, the average word segment duration is 0.287s, for mobile recordings it is 0.296s. The average phoneme segment duration is 0.071s for high quality recordings and 0.073s for mobile phone.

Table 4 shows the segment durations by phoneme class.

<table>
<thead>
<tr>
<th>class</th>
<th>mobile duration</th>
<th>high quality duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>approximant</td>
<td>0.099</td>
<td>0.073</td>
</tr>
<tr>
<td>diphthong</td>
<td>0.142</td>
<td>0.125</td>
</tr>
<tr>
<td>nasal</td>
<td>0.058</td>
<td>0.068</td>
</tr>
<tr>
<td>plosive</td>
<td>0.064</td>
<td>0.054</td>
</tr>
<tr>
<td>fricative</td>
<td>0.039</td>
<td>0.068</td>
</tr>
<tr>
<td>vowel</td>
<td>0.088</td>
<td>0.078</td>
</tr>
</tbody>
</table>

Table 4: Segment durations (in ms) by phoneme class for the read sentences

All phoneme classes are affected – there is a significant dependency between duration and signal quality ($F = 7.8768$, $p = 0.005$). Clearly, the fricatives show the strongest effect (see Figure 2). Here, the average phoneme duration for the mobile phone recording is only 57.3\% of that of the high quality recordings.

4. Discussion

The data presented here was computed by automated processes. The only difference between the high quality and the mobile phone recordings is the transmission channel and, subsequently, the signal quality. Any difference in the automatic labeling and segmentation of the signal must thus be due to this difference.

The MAUS system can be tuned to the different signal qualities by training the phoneme models, and by adapting weighting factors that govern the application of coarticulation rules. Hence, the results presented here are not a measure of the general performance of MAUS, but serve to illustrate the effects of different signal qualities.

4.1. Cutoff frequency

The different counts for fricative, plosive and vowel phonemes in the high quality and the mobile phone recordings may be attributed to the cutoff frequency in the mobile phone signal.

Fricatives, and the burst phase of plosives, have a large part of their energy above 4000 Hz, and this frequency range is not transmitted via the mobile phone or the ISDN channel. For an extreme example, see Figure 3, where the /s/ in /ʃɛnʃtʃ/ (window), clearly visible in the high quality signal, is totally missing from the mobile phone signal, thus yielding the segmentation /ʃɛntʃ/.

As a consequence, to the automatic segmentation algorithm of MAUS and in particular the phoneme models that were trained on high quality signals, these sounds are either missing in the signal and hence the segments are elided, or substituted by another phoneme, e.g. a substitution of a voiceless fricative by a voiced one.

A closer look at the segment counts reveals that the difference in vowel counts is almost entirely due to the /ɹ/. In high quality signals, the /ɹ/ is often elided through coarticulation, whereas in mobile phone quality signals MAUS with its standard settings does not apply this coarticulatory reduction very often.

An interesting detail is the fact that in mobile phone speech, the voiced plosives /b, d, g/ occur much more frequently than in high quality recordings (1600 vs. 1391 times), whereas the voiceless plosives are much more frequent in the high quality recordings (2024 vs. 1657). This may be due to the fact that the burst energy in voiceless plosives is lost in the mobile phone signal, leading to more plosives being classified as voiced in mobile phone speech.

The fricatives /h, v, x, s, f, z/ occur more often in high quality recordings, /ʃ/ occurs equally often in both recordings, and only /ʃ/ is more frequent in mobile phone recordings. Here, the effect of the cutoff frequency is especially clear – there are only a few traces of the fricatives left in the mobile phone signal, which leads to these segments being elided.

4.2. Durations

The differences in segment durations for high quality and mobile phone recordings affect mainly fricatives.

The duration of /ʃ, z, s, f, z/ for mobile phone is between 40.9\% and 62.3\% the duration of these phonemes in high quality recordings (and their counts differ between signal qualities).

If voiced and voiceless fricatives are viewed separately, it becomes clear that most voiceless fricatives in mobile phone signals have impossibly short durations, and that the voiced fricatives are only slightly longer (see Figure 4). A possible explanation for the short duration of fricative segments is that...
Figure 3: Signal fragment corresponding to the word *Fenster* in the high quality a) and the mobile phone b) signal. Note that the cutoff frequency in the mobile phone signal almost completely removes the phoneme /s/ from the mobile phone signal since almost all traces of friction in the mobile phone signal are filtered out, but no matching coarticulation rule for the phoneme can be applied, MAUS computes a minimally short fricative segment with approx. 20ms length. Voiced fricatives yield longer segments for mobile phone recordings (but still significantly shorter than for high quality signals); here MAUS finds traces of the fricative in the lower part of the spectrum and thus computes longer segments.

![Signal fragment](image)

Figure 4: Durations of voiced (VD) and voiceless (VL) fricative segments for high quality and mobile recordings

The segment durations of the front vowels /e, e:, i/ also are much shorter for mobile phone recordings than for their high quality recordings counterparts (70.2, 77.7 and 79.2%), but their counts do not differ very much. The other front vowels, e.g. /i/, /e/ are almost equal in duration for both recording qualities; in general, the further back and low a vowel, the longer its duration is in mobile phone recordings, e.g. for /o:/, /o/ the duration of the mobile phone segments is 125.9, 133.2 and 155.4% the length of its high quality counterpart.

5. Conclusion and outlook

This acoustical analysis of the difference in the automatic segmentation and labeling of mobile phone and high quality recordings has shown that both labeling and segmentation are affected. The effects are not uniform across all phonemes, not even within phoneme classes. Fricatives are the most affected phonemes, and the most consistent effect is the shortening of their segment duration and their reduced segment count in mobile phone speech. Within plosives, voiced and voiceless plosives differ in their effects on segment counts and durations.

Consonants in general, and fricatives in particular are very important as acoustic features in forensic phonetics as they have high perceptual confusability between speakers. Our results show that in the mobile phone signals fricatives are almost totally missing; this confirms findings of [10] who showed that the reduced signal quality of mobile telephone speech negatively affects speaker identification. Their analysis focused on spectral features of nasals; our analysis shows that also features such as segment counts and durations for nasals are significantly affected by the transmission channel.

The comparison of mobile phone and high quality recordings is a real world application in forensics. Here, quite often an original recording, in general via a fixed network or mobile phone, is available, and must be compared with high quality recordings of subjects during interrogation. In such an application, it is important to know what effects the signal quality may have on automated processes such as the MAUS system.

The present analysis is restricted in terms of speakers. Currently, further recordings are being performed at the Phonetics Institute within the ‘same-sex sibling’ comparison project by the second author. A further limitation, which is quite common for large speech databases, is that a manual verification of the results is in general not feasible because of time and budget constraints. Novel approaches to the visualization of results, e.g. an interactive browser for large speech databases, may alleviate this problem in the future.
6. References


[9] “clarin.phonetik.uni-muenchen.de/BASWebServices/.”