INDEPENDENT AUTOMATIC SEGMENTATION OF SPEECH BY PRONUNCIATION MODELING

Nicole Beringer, Florian Schiel
Department of Phonetics, University of Munich

ABSTRACT

In this paper we present an iterative automatic segmentation system which does not require any domain dependent training data. Input to the system is the canonical pronunciation and the speech signal of an utterance to be segmented, as well as a set of phonological pronunciation rules. The output is a string of phonetic labels (SAM−PA[1]) and the corresponding segment boundaries of the speech signal.

The system consists of three main parts:

In a first stage a set of general phonological rules is applied to the canonical pronunciation of an utterance yielding a graph that contains the canonic form and presumed variations.

In a second HMM−based stage the speech signal of the concerning utterance is time−aligned to this graph using a Viterbi search. The outcome of this stage is the time−aligned transcription of the input utterance.

Using this "raw" application of the phonological rules as the baseline in a third stage, a new set of statistically weighted rules is derived.

The procedure is repeated iteratively until the segmentation is not changed anymore.

1. INTRODUCTION

For many applications in speech processing, such as in ASR and speech synthesis (e.g. PSOLA), reliable segmentation and labeling of large speech databases is required. Also as ASR increasingly uses pronunciation modeling [2], [3] the demand for statistically based pronunciation models in different languages is growing.

Manual segmentation, especially for today's large speech corpora, is extremely time−consuming and a uniquely correct segmentation and labeling of an utterance does not exist because no two human experts are likely to produce exactly the same segmentation for the same utterance. Not even the same trained person will come to exactly the same transcription if asked to repeat the segmentation of the same utterance [4].

In previous work it was shown that human labelers can reach a correspondence of about 93.6% whereas the Munich Automatic Segmentation system − MAUS([5]) using domain−dependent pronunciation rules reaches 87.9% compared to human labellers on the German PHONDAT II corpus (read speech). A similar evaluation using spontaneous speech from the German VERBMOBIL project [6] resulted in 80.4% (human vs. human) and 78.5% (human vs. MAUS) respectively.

To produce competitive results, automatic segmentation methods like MAUS require a subset of manually labeled data (at least 1 h) from the domain of the corpus.

Since in most cases such data are not available we wish to become independent of a specific domain in order to avoid even partial manual segmentation of a corpus.

In this paper we present a system which is independent of a specific domain, i.e. which can be used for no matter what speech corpora as long as there exists a set of general phonetic rules like [7] for the corresponding language.

Section 2 describes the method of our iterative segmentation system. Section 3 shows the evaluation of this system by comparing data segmented manually and by MAUS.

Finally, results and future work are discussed in the last section.

2. ITERATIVE SEGMENTATION SYSTEM

For the following experiments we used a subset of the German part of the VERBMOBIL I corpus. The subset (approx. 5h of speech) was manually transcribed by skilled phoneticians and can be used as a reference for automatic segmentation.

Former work [5], [6] with the MAUS−system has shown, that the modeling of pronunciation variants represented by SAM−PA units and weighted with a−posteriori probabilities can be used successfully for the automatic segmentation and labelling of spontaneous German. MAUS uses a constraint search space derived from the canonical pronunciation of the given utterance in standard Viterbi alignment to come up with a broad phonetic transcript and a segmentation of the speech wave form. To compute the constraints for the search MAUS uses a set of statistical data−driven re−write rules that are automatically learned from approx. 1h of hand segmented data. In consequence, the rule−set is domain dependent and cannot be used effectively for other corpora.

The following examples show the form of these pronunciation rules:

```
I,n,#>I,# −5.565513
l,@,n,#>l,# −3.000183
n,Q,a>n,a −0.415490
g,@,n,t>g,N,t −1.641282
k,@,#>k,# −3.082453
aI,n,m>aI,m,m −1.392558
a:,Q,a:>a:,a: −0.020101
p,@,n,#>p,m,# −1.406395
P6,Q,O>P6,O −0.020101
N,@,n,#>N,N,# −0.275463
m,@,n,#>m,m,# −0.948678
```

The first label gives the left context, the last label before the ‘‘>’’ the right one. Between left and right context are the label(s) which change. On the right of the ‘‘>’’ the changed label(s) in the given contexts are noted. The last item shows the negative logarithmic probability of the occurrence of the

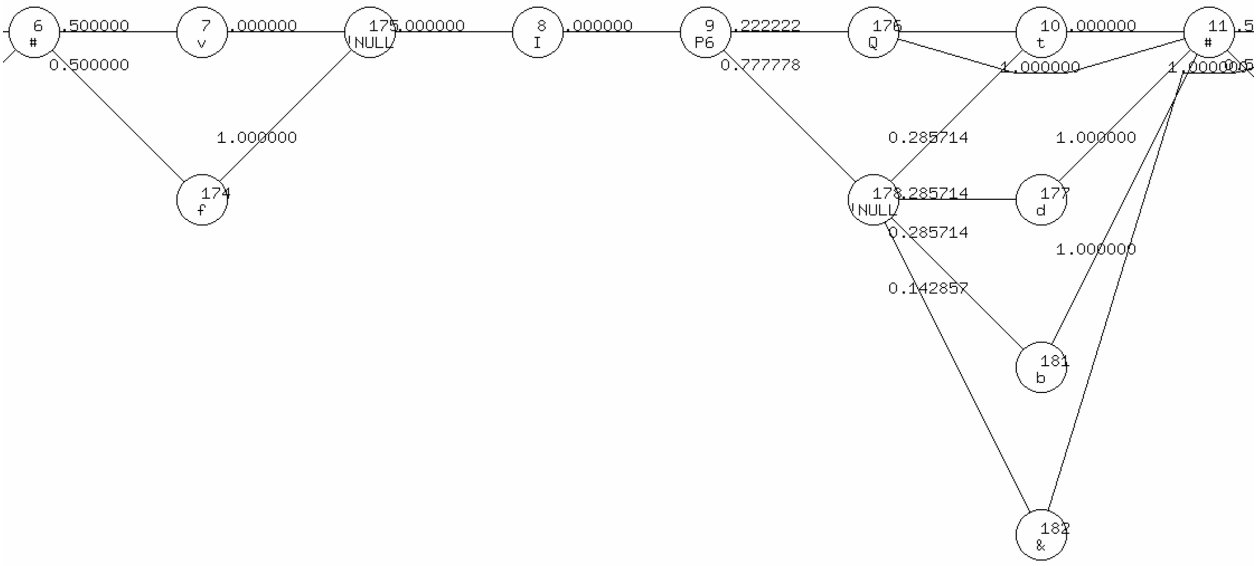

Figure 1. The MAUS pronunciation graph (rule iteration number 10) of the German word 'wird' (will be)

rule noted before.

In the work presented in this paper, we are integrating the MAUS principle for the automatic segmentation together with a learning algorithm for the rule set in an iterative process.

To start the iteration we used a fixed set of approx. 1500 general phonological rules of German pronunciation ([7]) compiled by an expert where every rule was assumed to be equally likely. It has been shown that such rules formulated a−priori are not effective for automatic segmentation, because too many unlikely pronunciation hypotheses are generated, which massively inflates the search space[5].

This rule set was used in a MAUS segmentation in order to obtain a first set M of pronunciation variants. M is compared to the canonical transcription of the corpus in order to statistically weight the applied rules during the segmentation. For instance, if words containing the final syllable /b@n/ in their citation form, e. g. »haben«, have been segmented in 80% of their occurances as [bm], then the rule b@n_> bm will get a higher statistical weight than a rule that had never any effect in the segmentation. A detailed description of the algorithm to calculate rules and their a posteriori probability can be found in [8].

It is important to note here that the original rule set is not altered by its contents but the a−posteriori probabilities for each observed rule are updated. Unseen rules remain equally likely on a lower level to ensure that new rules may be observed in a later iteration.

The whole process is repeated iteratively until we are sure that no changes in transcript and segmentation will occur.

Figure 1 shows a detail from the MAUS pronunciation graph after the tenth iteration for the utterance »ich will«. Nodes represent SAM−PA segments [1], while arcs give the transition probabilities between segments. Note that these probabilities are not the a−posterioris of the underlying rule set. For details on how to compute the graph from the rule set refer to [5]. In this example the MAUS segmentation resulted in the reduction of the last plosive [vI6] although this is not the most likely path in the graph.

```
Initialze rule set to general phonological rules
for k = 1 ... 10
    • MAUS segmentation yields transcript
    • computation of a−posteriori probabilities
        for each observed rule in the transcript
    • update of the general phonological rules
        due to computated probabilities
    • evaluation of the MAUS segmentations
        compared to reference transcript.
end for
```

Table 1: Algorithm for rule iteration

3. EVALUATION

3.1 Procedure

For the evaluation of these iteratively obtained data every segmentation of our test data (spontaneous speech) was compared to the hand segmented reference material. All labels of the corresponding utterances were matched to the reference labels by maximizing the number of identical label pairs. The matching labels were additionally evaluated in terms of segmentation boundaries. To simplify the evaluation we first only look at the quality of transcription and then for the quality of the corresponding segment boundaries.

3.2 Labeling correspondence

Table 2 shows the correspondence of the reference transcriptions and the transcriptions computed in the different iterations. Column one shows the phoneme error rate (PER) after the MAUS segmentation using the non weighted phonological rule set (called baseline). The PER is defined as the sum of replacements, deletions and insertions divided by the total number of labels in the reference transcriptions (2854). As we expected the PER is somewhere in the range of 27% (compared to earlier results by A. Kipp in [8]: 24.6%). Column 2 shows the PER after the first iteration, columns 3 and 4 the PER after the second iteration and the rest of the iterations respectively.

	Baseline	***Iteration1***	***Iteration2***	***Iteration3−9***
PER in %	26.80%	27.26%	23.72%	23.72%

Table 2: Percentage of the phoneme error rates among manual transcriptions and iterative automatic transcriptions.

It can be seen that the iteration process converges after the second iteration; no changes in terms of PER can be observed in higher iterations. Although there are no changes concerning the PER we found minor shifts in the segment boundaries between iteration 2 and 3. We obtained best results after the second iteration step (23.72%) compared to 21.5% using the domain−dependent rule set [8]. The increase in the PER after the first iteration (27.26%) was not expected. It is an open question whether this effect was due to the data set used or can be repeated on other corpora. We will verify this observation in future work with the RVG corpus ([9]).

The most frequent errors concerned the sound class of stops. Similar results were found in previous work [6].

3.3 Segment boundaries

The correspondence of the reference segmentations and the iterative automatic segmentations in terms of segment boundaries can be seen in the figures 2 to 5 for the corresponding cases of table 2. These show a histogram of the deviation of the automatic segmentation compared to the reference segmentation for matching transcripts. In this paper only the left boundaries are taken into account. All deviations bigger than |800| samples (±50 ms) were clustered in the bins on the edges of the range. Note that the speech signal has a sampling frequency of 16kHz.

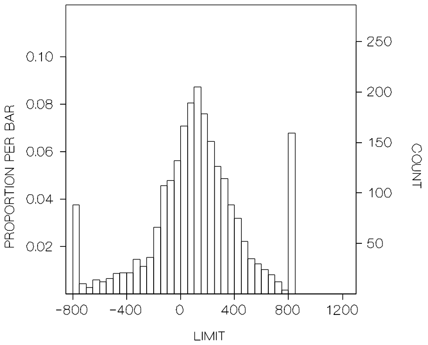

Figure 2: Distributions of relative frequencies of boundary deviation between the reference data and baseline

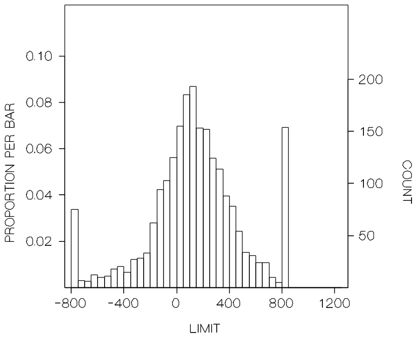

Figure 3: Distributions of relative frequencies of boundary deviation after iteration step 1

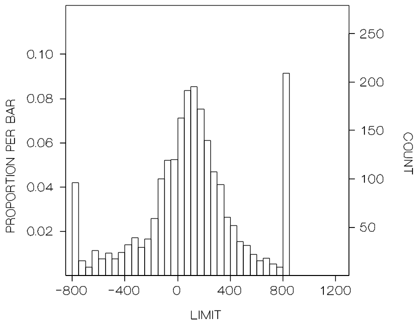

Figure 4: Distributions of relative frequencies of boundary deviation after iteration step 2

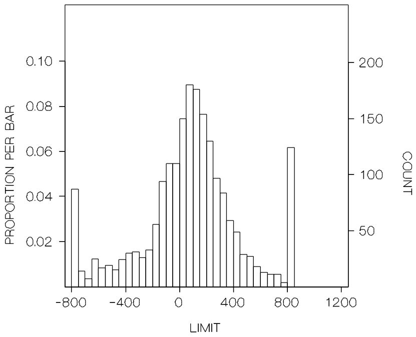

Figure 5: Distributions of relative frequencies of boundary deviation after iteration steps 3 − 9

The majority of segment boundaries (approx. 68%) lie within a window of ± 20 ms. However, it has to be noted that the distribution is not symmetric to the origin, but is shifted by approx. 6 − 7 ms. This has previously been observed in earlier MAUS experiments (e. g. [5]), and even by using a different HMM algorithm at a different lab. Our hypothesis is that this shift is due to an inherent processing problem within Hidden Markov Modelling.

4.DISCUSSION AND FUTURE WORK

The results show that it is possible to obtain high quality segmentation of speech signals by using iteratively weighted pronunciation rules. Although the iterative approach is not as good as the domain−dependent segmentation yet, we have non the less every confidence that we will be able to improve our results by using a bigger training set and by refining the alignment stage concerning the segment boundaries. It should also be taken into account that we may need a more detailed set of general pronunciation rules to statistically prune them down to the most relevant rules. The system is currently being revised by developing an algorithm to improve segment boundaries as well as by extending the rule contexts. As with all statistically based methods it is to be expected that results will improve proportionally to the amount of available data. Therefore one of our main efforts will be the production of MAUS segmented corpora.

REFERENCES

[1] http://www.phon.ucl.ac.uk/home/sampa/home.htm

[2] Ma, K, Zavaliagkos, G., Iyer, R.: Pronunciation Modeling for Large Vocabulary Conversational Speech Recognition ; in: Proceedings of the ICSLP 1998. Sydney, Vol. 6, pp. 2455−2458, Dec. 1998.

[3] Nicole Beringer, Florian Schiel, Peter Regel−Brietzmann (1998): German Regional Variants − A Problem for Automatic Speech Recognition?; in: Proceedings of the ICSLP 1998. Sydney, Vol. 2, pp. 85−88, Dec. 1998.

[4] B. Eisen, H. G. Tillman, and C. Draxler: Consistency of judgements in manual labeling of phonetic segments: The distinction between clear and unclear cases, Proc of the ICSLP (Banff), 1992, pp. 871−−874.

[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]M.−B. Wesenick, A. Kipp:Estimating the quality of phonetic transcriptions an segmentations of speech signals, Proc. of the ICSLP (Philadelphia), 1996.

[7] Maria−Barbara Wesenick (1996): Automatic Generation of German Pronunciation Variants; in: Proceedings of the ICSLP 1996. Philadelphia, pp. 125−128, Oct 1996.

[8] Andreas Kipp, Maria−Barbara Wesenick, Florian Schiel (1997): Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech. in: Proceedings of the EUROSPEECH, Sept 1997, Rhodos, Greece, pp. 1023−1026.

[9] Susanne Burger, Florian Schiel (1998): RVG 1 − A Database for Regional Variants of Contemporary German. in: Proceedings of the FIRST INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION 1998, Granada, Spain, pp. 1083−1087.