APPLYING SPEECH VERIFICATION TO A LARGE DATA BASE OF GERMAN TO OBTAIN A STATISTICAL SURVEY ABOUT RULES OF PRONUNCIATION

Maria-Barbara Wesenick, Florian Schiel [1]

Institut für Phonetik und sprachliche Kommunikation (IPSK)
Universität München, wesenick|schiel@sun1.phonetik.uni-muenchen.de

ABSTRACT

In this paper we present a new research project to obtain a statistical survey of the pronunciation of German using an automatic system for segmentation and labeling of speech data and a very large data base of spoken German (GermAn Spoken in Public, GASP). It mainly involves the development of two components: a) An automatic system of speech verification (PHONSEG) which produces a segmentation with semi-continuous HMMs corresponding to a given input-string of phonetic segment labels. b) A rule corpus of German pronunciation (PHONHYP) with which various possible forms of pronunciation can be derived from a citation form in order to model the variability of speech on a relatively broad segmental level. The rules express processes in German that are well-known and that can be observed in manual transcriptions as for example contained in the German PhonDat database.

The input to the system as a whole is the speech wave and the orthographic representation or the citation forms of the words contained in the utterance as a string of phonetic symbols. PHONSEG matches all possible forms that have been derived by PHONHYP on the signal and evaluates the likelihood. The output is the segmentation with the highest maximum overall likelihood, the corresponding

[1] currently at International Computer Science Institute, Berkeley

labels and for each segment the time-normalized log likelihood.

I. INTRODUCTION

For many speech processing tasks a representation of the pronunciation of the language concerned is required and this is usually taken from common pronunciation dictionaries. This is problematic for a number of reasons. The main problem however is, that dictionaries often have the aim to give the "right" or the "best" pronunciation of a language and therefore mostly give only one possible form which is usually not the most common form of pronunciation. But in spoken language there is not just one possible way to pronounce a word. In fact, no two utterances are ever produced exactly the same.

As is known the variability of utterances has different causes: it is due to regional or dialectal differences among speakers, to their social background, to the speaking style and to the kind of communication channel that is used. Other factors are age, gender and physiology of the vocal tract of the speaker. Therefore, in speech technology concrete, complex and consistent information about possible variation in pronunciation is required. Especially in segment-based speech recognition applications it is indispensable to process as much information about variation in pronunciation as possible to be able to analyze the multifold input of human speech. Therefore we need fundamental knowledge of what phenomena may be observed in spoken language to be able to develop reliable systems in the area of speech technology.

Although pronunciation processes in German are well known and have often been described a broad empirical investigation about the occurrence of word forms has not yet been carried out. To obtain enough information about phonetic processes as a basis for statistical investigations that actually can occur it is essential to evaluate large databases of continuous speech. For the evaluation on the segmental level the availability of segmentation and labeling of the speech material is fundamental. This may also be useful for many speech processing tasks for example for training phoneme-based speech-recognizers. Manual labeling of speech material that is contained in very large databases is practically not possible for it would be very time-consuming. Furthermore, the segmentations would be subjective (however correct they were) and nevertheless prone to inconsistency and errors. To obtain fundamental knowledge and deep insights of what happens in spoken language, which phonetic processes take place and which variations are possible to which degree in which context etc. we need the help of reliable speech technology.

In this contribution we present a new project to obtain a statistical survey about the different possible forms of pronunciation of German words using an automatic system of speech verification (PHONSEG), a rule-corpus of German pronunciation with an algorithm for the creation of pronunciation hypotheses (PHONHYP) and a very large database of spoken German (GermAn Spoken in

Public, GASP). The project can roughly be divided into two parts:
a) The development of an automatic system for segmentation of the speech signal according to a given string of phonetic symbols. The output of this segmentation consists of the boundaries of the speech segments corresponding to an input-symbol and for each segment the time-normalized log likelihood (speech verification). The first stage of this system produces a rough segmentation with semi-continuous HMMs; at the second stage this segmentation is refined by a rule-based system.
b) The development of a rule corpus, with which possible variations of a citation form can be derived at the level of a relatively broad phonetic transcription. The rules express on the segment level coarticulation processes and other phonetic phenomena of German as have been observed by the analysis of manual transcriptions a great part of which are contained in the German PhonDat database. Most of these phenomena are well known and have often been described in the literature (e.g. [5],[6]).

Combining these two systems phonetic/phonemic investigations on German can be carried out using GASP as follows:
Using the rule corpus PHONHYP on the symbolic level all possible variations of the sentence citation form of each utterance in GASP are produced. By using the speech verification system PHONSEG it is possible to decide which of these variations represent the actual utterance best. By analyzing this output of the system (including segmental information, transcriptions, rules applied) we obtain an empirically based survey about the phonetic-phonological variations of pronunciation in the database. In the course of the development, by evaluating the system performance and by the analysis of occuring errors iteratively improvements can be made on both system components. The obtained knowledge will certainly be useful not only for speech technology but also for phonologists and phoneticians.

The following sections describe very briefly the automatic speech verification system PHONSEG and the generation of possible pronunciation forms of the sentences that are contained in GASP using PHONHYP. Section IV gives a brief survey of the ongoing work to build up the GASP database. Sections V and VI list some of the aims of the whole project regarding the investigation of phonetic-phonologic processes and a discussion about the first results using the described system.

II. SPEECH VERIFICATION

Figure **??** gives an overview of the automatic speech verification system. The PHONSEG system accepts the speech wave of the utterance and a string of phonetic-phonologic symbols (hypothesis) as input. It produces the following output:

- beginning and ending of the utterance within the signal

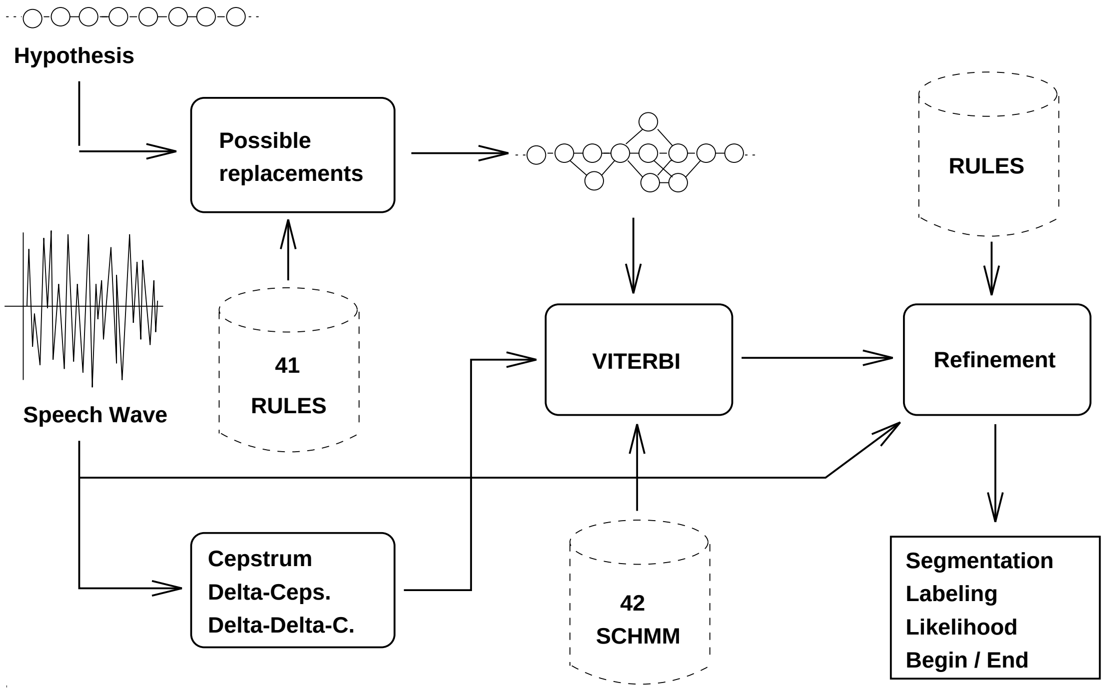

Figure 1: Automatic speech verification PHONSEG

- labeling and segmentation oriented by the input string
- for each segment the time-normalized log likelihood

In the first stage PHONSEG runs a constrained Viterbi over the input string (augmented by leading and trailing silence) using context-free semi-continuous HMMs (scHMM) modeling 42 phoneme classes (extended SAM-PA):

- Preprocessing into 12 cepstral coefficients + energy + zero crossing rate + 1st and 2nd derivative every 10 msec
- 5 codebooks with diagonalized covariance matrices
- 3 to 6 states per scHMM; product code
- Initialization with data segmented by hand (approx. 2400 utterances from 12 speakers)
- Training with segmental-k-means ([4]) on the PhonDat 2 corpus (16110 utterances from 201 speakers).

During the forced Viterbi search replacements according to the following rules are allowed:

- 27 rules regarding replacement of certain vowels in any context

- replacement of 7 voiced consonants into voiceless consonants when in voiceless left or right context
- replacement of 7 voiceless consonants into voiced consonant when in voiced left or right context

The labeling and segmentation from the Viterbi search is passed to the second stage of PHONSEG which refines many boundaries according to a small rule system (note that the labeling is in most cases not the same as the input string, because of possible replacements). Without going into details this rule system handles the following problems:

- merge certain vowel combinations to diphthongal segments
- correct crucial transitions from voiced consonants to vowels or diphthongs and vice versa
- correct mostly wrong left boundaries of plosives and aspirants as initial segments
- finally shift calculated boundaries into the next positive zero crossings

From the resulting segmentation and the backtracking of the Viterbi search we calculate the time-normalized log likelihood ('error measure') of each segment. To explore potential dependencies between the error measure values together with the segment durations computed by the automatic segmentation system and the consistency classes the following comparisons were carried out. 64 sentences (read by six speakers each) from the train enquiries corpus of the PhonDat 2 database were manually segmented and labeled by at least four trained human segmenters. These segmentations are now being compared to the one produced by the first version of the automatic segmentation system.
Three segment label consistency classes were defined:
(I) agreement of manual and automatic labeling
(II) agreement of manual segmentation but disagreement with automatic labeling
(III) disagreement within the manual and automatic labeling
First results suggest that the elision of the schwa, a common reduction phenomenon in spoken German, correlates with a particularly low negative error measure value. Hence, a low negative error measure value, together with a short segment duration, is an indication that this phoneme is not produced in the current utterance. A subsequent re-mapping of the citation form with the phoneme removed and the speech wave results in better error measure values for the other segment labels.

For a full report on the statistical comparisons see [2].

III. GENERATION OF HYPOTHESES

The systems component PHONHYP has been designed out of the necessity for a transcription as input to PHONSEG together with the speech signal. In an earlier version of PHONSEG the input string of phonetic-phonologic symbols were just the concatenated citation forms of the words in the uttered sentence. Obviously the error rate was particularly high in cases when the segmental structure of the actual utterance had changed compared to the input citation form due to phonetic processes such as elisions or assimilations. The idea to improve the system was to offer alternative pronunciation forms together with the citation forms and to let the system evaluate which of the forms matches best with the signal. We decided to produce these forms by a set of rules that express possible segmental alterations of the citation form. To get a solid basis of rules the manual transcriptions of the PhonDat 2 corpus of spoken German which is stored in a database component of the Prolog environment eclipse have been analyzed. This is possible in an efficient and systematic way by using Prolog tools ([1]). Also reduction-phenomena that are well-known and have been described in literature and a number of hypothesized changes have been put in rules. The rule corpus has been growing during our work up to a number of approx. 1700 rules of segmental changes. (See [8] for a detailed description of PHONHYP.)

The algorithm that applies the rules on the citation form can be very briefly explained as follows:

```
Build ´sentence citation form´ of the utterance
Mark all rules of the corpus as ´usable´
Do until no new variation is produced
    Do for all rules marked as ´usable´
       Test if rule can be applied to input
       (citation form of the utterance)
       If yes
          Produce ONLY NEW variations of input
          AND of all variations produced so far
       Else
          Mark rule as ´not usable´
```

The citation forms of the words are derived of a large pronunciation dictionary of German (approx. 90000 lemmata) developed and maintained by D. Stock, University of Bonn, Germany.

We chose this slightly complicated strategy because of two reasons: on the one hand the order of the rules should have no influence on the output of PHONYP; on the other hand to prevent a fully recursive generation, which would result in a huge number of outputs most of which would not make sense. As a tradeoff we decided to design the rules in such a way that they have to be applied in a first step to the citation forms only. Then only the applicable rules are applied recursively to citation forms and new variants until no new form can be derived.

Orthographic form:		Regensburg	
Sentence citation form:		re:g@nsbU6k	
rules that can be applied:			
1g@n>gn		(Elision of schwa)	
1g@n>gN		(assimilation of place)	
1gN>N		(assimilation of manner)	
1ns>nz		(assimilation of voice)	
1sb>sp		(assimilation of voice)	
1ns>nts		(consonant epenthesis)	
1sb>zb		(assimilation of voice)	
Resulting forms:	re:gNzbU6k	re:gNsbU6k	re:NsbU6k
re:g@nzbU6k	re:gnzbU6k	re:g@nspU6k	re:gnspU6k
re:gNspU6k	re:NspU6k	re:g@ntsbU6k	re:gntsbU6k
re:g@ntspU6k	re:gnsbU6k	re:g@ntzbU6k	re:gntzbU6k
re:gntspU6k	re:NzbU6k	re:gNtsbU6k	re:gNtsbU6k
re:NtsbU6k	re:NtsbU6k	re:gNtzbU6k	re:NtzbU6k

Table 1: Generation of pronunciation forms by PHONHYP (extended SAM-PA)

Some examples for rules and the generation of variants from a citation form are shown in table **??**

IV. GASP – GermAn Spoken in Public

In close cooperation with the Institut der deutschen Sprache in Mannheim (IDS) and the Leibniz Rechenzentrum in Munich (LRZ) we are currently producing a database GASP that will cover a wide range of spoken German; this includes read speech, standard German (for instance radio news), talks, interviews and free dialogues. GASP will contain only the speech wave and an orthographic transcription of the recorded speech; the length of the single items is not limited. Main sources for GASP are recordings done at our site and recordings from the last decades done at the IDS. The latter are recordings done in analog technique and are currently digitized and prepared for the corpus at our site. In the first preparatory phase we will gather approx. 150 hours of speech in a standard format of 16 bit/16 kHz. The problems of memory and data access of such a huge database is addresses by a common project with the LRZ.

V. PHONETIC/PHONOLOGIC PROCESSES

By combining the above described system PHONSEG and PHONHYP we will perform an investigation of the GASP database. The rough procedure for each item in GASP is: look up the orthographic words in a citation form dictionary and build up a 'sentence citation form' for the whole utterance, generate all posssible

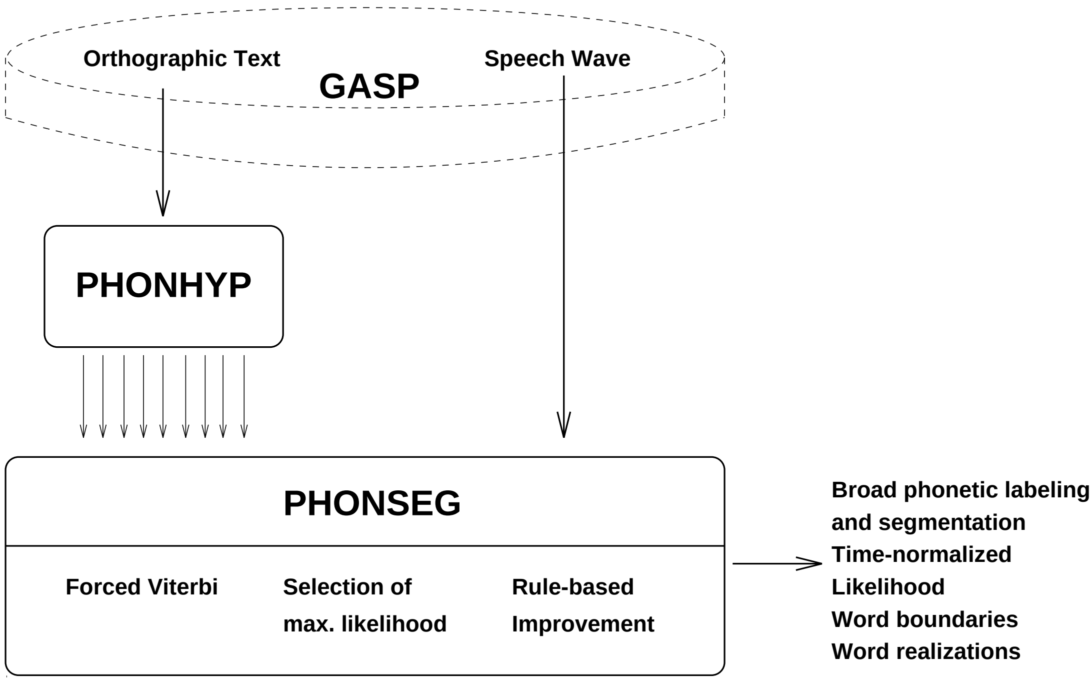

Figure 2: Processing the GASP database using PHONHYP and PHONSEG

variations of the sentence citation form and represent them in a suitable form by using PHONHYP, select the most likely hypothesis and calculate the output mentioned in section II by using PHONSEG. Figure **??** shows the combined system. By analyzing the output we expect the following valuable results for the different areas of interest:

- Phonetics: sufficient and consistently labeled and segmented speech data to investigate phonetic processes on a segmental level, the occurrence of reduction-phenomena, the effects at word boundaries
- Phonology: significant statistics about word realizations or the usage of pronunciation rules, verification of phonotactical or phonological models
- Speech Technology: reliable pronunciation models, consistently labeled and segmented data for speech recognition and speech synthesis, markers for events in speech, where traditional modeling is highly probable to fail

At the current stage the system is capable to process read utterances up to 20 sec. The computation time depends heaviliy upon the numbers of possible pronunciations to verify. The average computation time on a Sparc 10 is currently 800 sec per utterance. At the moment the different possible pronunciations are represented by a simple list, which is of course not very efficient.

VI. PRELIMINARY RESULTS

1	Das ist nun doch die Hoeh!
	das QIst nu:n dOx di: h'2:!
	daz Is nu:n dOx di: h'2:!
2	Motoren brauchen Benzin, Oel und Wasser.
	mo:t'o:r@n br'aUx@n bEnts'i:n, Q'2:l QUnt v'as6.
	mo:t'o:r@m br'aUxN bEnts'i:n, Q'2:l Un f'as6.
3	Die drei Maenner sind begeistert.
	di: dr'aI m'En6 zInt b@g'aIst6t.
	di: dr'aI m'En6 zInp b@g'aIs6t.
4	Stehend macht man seine Aussage.
	St'e:h@nt m'axt man zaIn@ Q'aUsza:g@.
	Sd'e:nb m'axp man zaIn@ Q'aUsa:g@.
5	Abends lieber zeitig schlafen gehen.
	Q'a:b@nts l'i:b6 ts'aItIC Sl'a:f@n g'e:h@n.
	Q'a:bms l'i:b6 ts'aItI Sl'a:f@n g'e:n.
6	Die Aerzte sind damit gar nicht einverstanden.
	di: Q'E6tst@ zInt da:mIt g'a:6 nICt Q'aInf6Sdand@n.
	di: Q'E6s@ zIn & da:mI g'a:6 nICt Q'aInf6Sdan.

Table 2: Orthographic representation, citation forms and automatic transcription of 6 utterances (extended SAM-PA, ' primary stress, Q glottal stop, & arbitrary word boundary)

So far we have not developed a method for an automatic evaluation of the output, which is especially difficult as we have no reference transcription to compare the results with. As it is widely known, segmentations by human experts are not consistent in particular classes of phonemes ([3]). At the moment the segmentation output is evaluated only qualitatively by analyzing it in form of time signal and sonagram which is sufficient at this stage to find the obvious weak points of the system.

As can be seen in the transcriptions of Table 2, a number of pronunciation forms that show segmental changes compared with the corresponding citation form have been chosen by the system as more likely to represent the speech signal than the citation form. The segmental changes include phenomena that are typical for spoken German as schwa-elision or assimilation of place, manner and voice. It can be said that the transcriptions are very promising, because they are plausible and very likely to be correct. It is indispensable to control the system's output auditively and visually by analyzing different representations (e.g. oscillogram, spectrogram) of the speech signal to be able to evaluate the performance of the system. The analysis of just a few segmentations show the following tendencies:

- most boundaries are correct with a maximal deviation of 10 - 15ms
- laryngalizations are well detected

- the beginning and end of utterances are often not correct (deviations up to 80ms)
- boundaries of plosives are often not precise enough (although within 10ms maximal deviation)
- vowel clusters and vowel-lateral combinations are not well detected
- labels of voiced and voiceless plosives are sometimes confused
- plosives with no complete closure are not detected
- the label for schwa is used too often and schwa-elisions are not always detected

VII. FUTURE WORK

The rule-based system in PHONSEG has to be improved to handle typical errors of the Viterbi alignment. Furthermore we will change the representation of the possible hypotheses generated by PHONHYP into a finite state machine and improve the data access via the Internet to speed up the search. Other problems are the treatment of very long signals (more than 200 sec), hesitations, non-speech phenomena and unknown words (names, locations, etc.). The rule corpus PHONHYP has to be evaluated to avoid unnecessary search. By the use of morphologic knowledge we hope to decrease the huge amount of possible transcriptions in longer utterances. Finally we will continuously extend the GASP database to a size that satisfies our statistical requirements.

Parts of this project were funded by the Bundesministerium für Forschung und Technologie, Bonn (contract number 413-4001-01 IV 102 L/4).

REFERENCES

[1] Ch. Draxler, H.G. Tillmann, B. Eisen: Prolog tools for Accessing the PhonDat Database of spoken German, Proceedings of the 3rd EUROSPEECH 1993, Berlin, Sept 1993.
[2] Ch. Draxler, H.G. Tillmann, F. Schiel: Manual and Automatic Segmentation of the PhonDat Database of Spoken German, Technical Report, Institut für Phonetik und sprachliche Kommunikation, Universität München, to appear.
[3] B. Eisen, H.G. Tillmann, Ch. Draxler: Consistency of judgements in manual labeling of phonetic segments: the distinction between clear and unclear cases, Proceedings of the ICSLP 1992, Banff Canada, pp. 871 - 874, Nov 1993.
[4] X.D. Huang, M.A. Jack: Hidden Markov Modelling of speech based on semi-continuous model, Electronics Letters, Vol. 24, No. 1, pp. 6 - 7, 7th Jan 1988.
[5] K.J. Kohler: Segmental Reduction in Connected Speech in German: Phonological Facts and Phonetic Explanations, in: W. J. Hardcastle, A. Marchal (eds.):

Speech Production and speech Modelling. Kluwer, pp 69 - 92, 1990.
[6] G. Meinholt, E. Stock: Phonologie der deutschen Gegenwartssprache, Leipzig, 1982.
[7] H.G. Tillmann, B. Pompino-Marschall: Theoretical Principles Concerning Segmentation, Labelling Strategies and Levels of Categorical Annotation for Spoken Language Database Systems, Proceedings of the 3rd EUROSPEECH 1993, Berlin, pp. 1691 - 1694, Sept 1993.
[8] M.-B. Wesenick: Entwurf eines Regelsystems der Aussprache des Deutschen als Basis fuer empirische Untersuchungen. M.A. thesis, IPSK Ludwig-Maximilians-Universität München 1994.