PROBABILISTIC ANALYSIS OF PRONUNCIATION WITH 'MAUS'

Florian Schiel, Andreas Kipp

Institut für Phonetik und Sprachliche Kommunikation, Ludwig-Maximilians-Universität München

ABSTRACT

This paper describes a method to automatically detect pronunciation variants in large speech corpora within the framework of the 'MAUS' project ([1]). 'MAUS' stands for 'Munich Automatic Segmentation System' and is a general purpose tool to automatically label and segment read and spontaneous German speech into phonetic/phonologic segments. The output of MAUS can for example be used to build probabilistic models of pronunciation of fluent German as reflected by the analysed corpus. These models can be the basis for phonetic investigations or can be incorporated into classic speech recognition algorithms.

The paper is organised as follows: The first section gives a very short introduction into the main processing principle of MAUS and gives some examples of the output of MAUS applied to utterances from the Verbmobil corpus. Section 2 deals very briefly with the problem of how to evaluate such an output. A method is given that first compares the performance of three human transcribers with each other and then the performance of MAUS with each of them. Section 3 describes our method for deriving probabilistic pronunciation dictionaries from the MAUS output and gives some interesting examples from the Verbmobil domain. The 4th and last section gives some new approaches towards incorporating these models into a new automatic speech recognition (ASR) approach that combines phonetically 'sharper' acoustic models with the probabilistic modelling of pronunciation.

1. INTRODUCTION TO MAUS

The MAUS system was developed at the Bavarian Archive for Speech Signals (BAS) to facilitate the otherwise very time-consuming manual labeling and segmentation of speech corpora into phonetic units. Initially funded by the German government within the Verbmobil I project, MAUS is now further extended by BAS with the aim to automatically improve all BAS speech corpora by means of complete broad phonetic transcriptions and segmentations. The basic motivation for MAUS is the hypothesis that automatic speech recognition (ASR) of conversational speech as well as high quality 'concept-to-speech' systems will require huge amounts of carefully labelled and segmented speech data for their successful progress.

Traditionally a small part of a speech corpus is transcribed and segmented by hand to yield bootstrap data for ASR or basic units for concatenative speech synthesis (e.g. PSOLA). Examples for such corpora are the PhonDat I and II corpus (read speech) and the Verbmobil corpus (spontaneous speech). However, since these labelings and segmentations are done manually, the required time is about 800 times the duration of the utterance itself, e.g. to label and segment an utterance of 10 sec length a skilled phonetician spends about 2 h and 13 min at the computer. It is clear that with such an enormous effort it is impossible to annotate large corpora like the Verbmobil corpus with over 33 h of speech. On the other hand large databases are needed urgently for empirical investigations on the phonological and lexical level.

Input to the MAUS system is the digitised speech wave and any kind of orthographic representation that reflects the chain of words in the utterance. Optionally there might be markers for non-speech events as well, but this is not essential for MAUS. The output of MAUS is a sequence of phonetic/phonemic symbols from the extended German SAM Phonetic Alphabet ([5]) together with the time position within the corresponding speech signal.

Example:

```
Input:
Speech Wave + ´bis morgen wiederhoeren´

Output:
MAU: 0 479 -1 <p:>
MAU: 480 480 0 b
MAU: 961 478 0 I
MAU: 1440 1758 0 s
MAU: 2720 959 1 m
MAU: 3680 799 1 O
MAU: 4480 2399 1 6
MAU: 6880 2079 1 N
MAU: 8960 799 2 v
MAU: 9760 959 2 i:
MAU: 10720 479 2 d
MAU: 11200 2239 2 6
MAU: 13440 799 2 h
MAU: 14240 639 2 2:
MAU: 14880 1439 2 6
MAU: 16320 1599 2 n
MAU: 17920 1759 -1 <p:>
```

(The output is written as a tier in the new BAS Partitur format. 'MAU:' is a label to identify the MAUS tier; the first integer gives the start of the segment in samples counted from the beginning of the utterance; the second integer gives the length of the segment in samples; the third number gives the word order and the final string is the labeling of the segment in extended German SAM-PA. See [10] for a detailed description of the BAS Partitur format)

MAUS is a three-staged system (see fig.1):

In a first step the orthographic string of the utterance is looked up in a canonical pronunciation dictionary (e.g. PHONOLEX, see [8]) and processed into a Markov chain (represented as a directed acyclic graph) containing all possible alternative pronunciations using either a set of data driven microrules or using the phonetic expert system PHONRUL.

A microrule set describes possible alterations of the canonical pronunciation within the context of $+/-1$ segments together with the probability of such a variant. The microrules are automatically derived from manually segmented parts of the corpus. Hence, these rules are corpus dependent and contain no a priori knowledge about German pronunciation. Depending on the pruning factor (very sel-

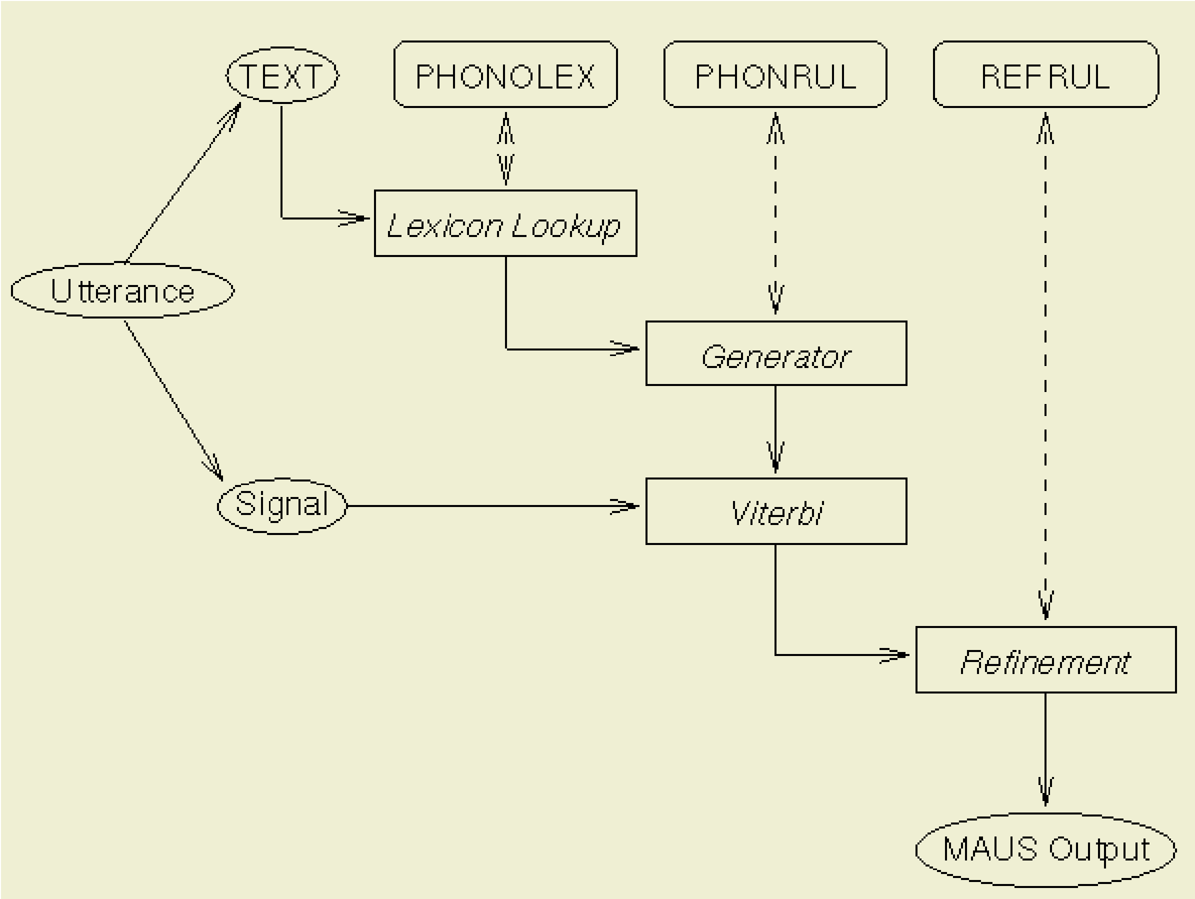

Figure 1: The MAUS system - block diagram

dom observations are discarded) and the size of the manually segmented data the microrule set consists of 500 to 2000 rules. In this paper we use a set of approx. 1200 rules derived from 72 manually segmented Verbmobil dialogs of The Kiel Corpus of spontaneous Speech ([6]). Details about this method can be found in [1].

The expert system PHONRUL consists of a rule set of over 6000 rules with unlimited context. The rules were compiled by an experienced phonetician on the basis of literature and generalised observations in manually transcribed data. There is no statistical information within this rule set; all rules are treated with equal probability. PHONRUL is therefore a generic model and should be considered independent of the analysed speech corpus. A more detailed description of PHONRUL can be found in [7].

The second stage of MAUS is a standard HMM Viterbi alignment where the search space is constrained by the directed acyclic graph from the first stage (see figure 2 for an example). Currently we use the HTK 2.0 as the aligner ([9]) with the following preprocessing: 12 MFCCs + log Energy, Delta, Delta-delta every 10 msec. Models are left-to-right, 3 to 5 states and 5 mixtures per state. No tying of parameters was applied to keep the model as sharp as possible. The models were trained to manually segmented speech only (no em-

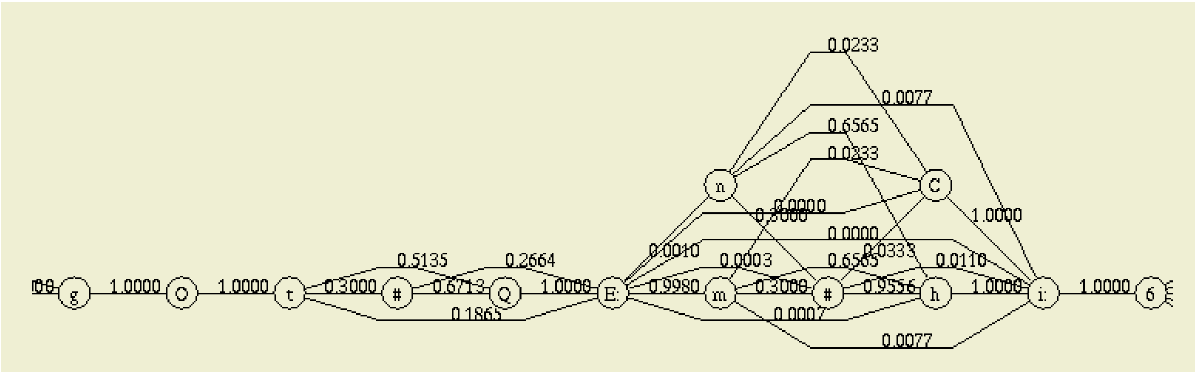

Figure 2: Acyclic graph of the utterance "Gott... ähm... hier..." with possible pronunciation variants

bedded re-estimation).

The outcome of the alignment is a transcript and a segmentation of 10 msec accuracy, which is quite broad. Therefore in a third stage REFINE the segmentation is refined by a rule-based system working on the speech wave as well as on other fine-grained features. However, the third stage cannot alter the transcript itself, only the individual segment boundaries.

The general drawback of the MAUS approach is, of course, that MAUS cannot detect variants that are not 'foreseen' by the first stage of the process. However, we found that using the microrule method the vast number of distinct rules are found after analyzing a relatively small subportion of the whole corpus. This indicates that the number of non-canonical pronunciations occurring in a certain domain such as the Verbmobil corpus is in fact limited and therefore treatable by a limited number of rules.

2. EVALUATION

The output of MAUS can be separated into two different classes: the transcript (the chain of symbols) and the corresponding segmental information (begin and end of each segment).

Unlike in an ASR task the evaluation of a phonetic/phonemic segmentation of arbitrary utterances has a great disadvantage: there is no reference. Even very experienced phoneticians will not produce the same segmentation, not even the same transcript on the same speech wave.

We tried to circumvent this general problem by first comparing the results of three experienced human transcribers on the same corpus with each other to get a feeling for what is possible and set an upper limit for MAUS. We used standard Dynamic Programming techniques as used in ASR evaluations (e.g. [9]) to calculate the inter-labeller agreement between different transcripts. We found that the coverage of the three human transcribers ranges from 78.8% to 82.6% (on the basis of approx. 5000 segments). We then calculated the accuracy for the MAUS output with regard to each set of human results and found values ranging from 74.9% to 80.3% using the microrule method and 72.5% to 77.2% using PHONRUL. Not surprisingly, the worst and best coverage were correlated in all three experiments. This means that if we set the upper limit to the best match within human transcription results (82.6%) and compare this to the average agreement of MAUS with these two human transcribers, we'll end up with a relative performance of 97.2% for MAUS. (Note that this relative performance measure might be higher than 100% at some distant point in the future!)

For a more detailed discussion about the problem of evaluation as well as a more accurate analysis of the MAUS output (applied to read speech) please refer to [3].

In terms of accuracy of segment boundaries the comparison between manual segmentations shows a high agreement: on average 93% of all corresponding segment boundaries deviate less than 20msec from each other. The average percentage of corresponding segment boundaries in a MAUS versus a manual segmentation is only 84%. This yields a relative performance of 90.3%. We hope that a further improvement of the third stage of MAUS will increase these already encouraging results.

3. PROBABILISTIC PRONUNCIATION MODEL

Aside from the many other uses of the MAUS output for this paper we'll show how to derive a simple but effective probabilistic pronunciation model for ASR from the data. There are two obvious ways to use the MAUS results for this purpose:

- use direct statistics of the observed variants
- use generalised statistics in form of microrules

In the following we will discuss both approaches.

3.1. Direct Statistics

Since in the MAUS output each segment is assigned to a word reference level (Partitur Format, see [10]), it is quite easy to derive all observed pronunciation variants from a corpus and collect them in a PHONOLEX ([8]) style dictionary. The analysis of the training set of the 1996 Verbmobil evaluation (volumes 1-5,7,12) led to a collection of approx. 230.000 observations.

The following shows a random excerpt of the resulting dictionary:

```
terminlich
adj
t E 6 m i: n l I C
t E 6 m i: n I C        3
t @ m i: l I C  3
t E 6 m i: n l I C      10
t E 6 m i: l I C        1
t @ m i: n l I C        7
&
...
Karfreitag
nou
k a: 6 f r aI t a: k
k a: 6 f r aI t a: k    15
k a: 6 f r aI t a x     3
&
...
weil
par
v aI l
v a l   11
v aI    108
v aI l  207
&
...
siebenundzwanzigsten
adj
z i: b @ n U n t t s v a n t s I C s t @ n
z i: b @ n U n s v a n t s I s t @ n    1
z i: b m U n s v a n t s I k s t n      2
z i: b m U n s v a n s I C s t n        1
z i: b m U n s v a n t s I C s t @ n    1
z i: m U n s v a n t s s t @ n  1
z i: m U n s v a n t s s n      1
z i: b m U n s v a n t s s t    1
s i: b @ n U n s v a n s I C s t n      1
```

```
z i: b @ n U n s v a n s I C s t n      1
z i: b @ n U n s v a n t s I s t n      3
z i: m U n s v a n t s I s t @ n        1
z i: b m U n s v a n s I z n    2
i: b m U n s v a n z I z n      1
z i: m U n s v a n t s I z n    6
z i: m U n s v a n t s I s n    1
z i: b m U n s v a s I s t n    1
z i: b @ n U n s v a n t s I C s t n    2
z i: m U n s v a n s I s n      1
z i: m U n s v a n z I k s t @ n        1
z i: b m U n s v a n t s I z n  2
z i: m U n s v a n t s I s t    2
z i: b m U n s v a n t s I s t @ n      2
z i: m U n s v a n s I s t n    17
z i: m U n s v a n s s t n      1
z i: b m U n s v a n s I C s t  1
z i: m U n s v a n t s I C s n  1
z i: b m U t s v a n t s s t n  1
z i: m U n s v a n s I k s t @ n        2
z i: b m U n s v a n s I s t n  6
z i: b m U n s v a n s I k s t n        1
z i: b m U n s v a n s I s t @ n        4
z i: m U n s v a n t s I k s t n        6
z i: m U n s v a n t s s t n    1
z i: m U n s v a n s I z n      1
z i: b m U n s v a n s s t n    2
z i: b m U t s v a n s I s t n  1
z i: m U n s v a s s n  1
z i: b @ n U n s v a n t s I k s t n    2
z i: m U n s v a n t s I C s t n        9
z i: m U n s v a n z I z n      2
z i: m U n s v a n t s I s t n  27
z i: b m U n s v a n t s s t n  1
z i: m U n s v a n z I s t n    5
z i: b m U n s v a n s I s n    1
z i: m U n s v a n s s n        1
z i: m U n s v a n s s t @ n    1
z i: b m U n s v a n z I z n    1
z i: m U n s v a n t s I C s t @ n      2
s i: b m U n s v a n t s I C s t n      1
z i: b m U n s v a n t s I C s t n      12
z i: m U n s v a n s s t        1
z i: m U n s v a n s I C s t n  3
z i: m U n s v a n z I s n      1
z i: b m U n s v a n t s I s t n        28
z i: m U n s v a n s I s t @ n  3
z i: b m U n s v a n t s I k s t @ n    1
&
...
Namen
nou
n a: m @ n
n a: m  30
n a: m @ n      15
&
...
Essen
nou
Q E s @ n
@ s n   2
E s n   16
E s @ n 6
s n     3
E s     1
Q E s @ n       7
Q E s   1
Q E s n 21
&
```

The above modified PHONOLEX format is defined as follows:

```
<orthography>
<comma separated list of
linguistic classes>
<canonical pronunciation>
<empiric pronunciation> <count>
&
...
```

Obviously many of the observations are not frequent enough for a statistical parameterisation. Therefore we prune the baseline dictionary in the following way:

- Observations with a total count of less than N per lexical item are discarded.
- From the remaining observations for each lexical word L the a-posteriori probabilities $P(V|L)$ that the variant V was observed are calculated. All variants that have less than M% of the total probability mass are discarded.
- The remaining variants are renormalised to a total probability mass of 1.0.

Applied to the above example this yields the following more compact statistics (pruning parameters: N=20, M=10):

```
terminlich            0.434783
t E 6 m i: n l I C
terminlich            0.130435
t E 6 m i: n I C
terminlich            0.304348
t @ m i: n l I C
terminlich            0.130435
t @ m i: l I
Karfreitag            1.000000
k a: 6 f r aI t a: k
weil                  0.342857
v aI
weil                  0.657143
v aI l
siebenundzwanzigsten 0.509091
z i: b m U n s v a n t s I s t n
siebenundzwanzigsten 0.490909
z i: m U n s v a n t s I s t n
Namen                 0.333333
n a: m @ n
Namen                 0.666667
n a: m
Essen                 0.320000
E s n
Essen                 0.420000
Q E s n
Essen                 0.120000
E s @ n
Essen                 0.140000
Q E s @ n
```

where the second column contains the a-posteriori probabilities. This form can be directly used in a standard ASR system with multi pronunciation dictionary like HTK (version 2.1).

3.2. Generalised statistics

The usage of direct statistics has the disadvantage that most of the words will be modelled by only one variant, which in many cases will be the canonical pronunciation because of lack of data. An easy way to generalise to less frequent words (or unseen words) is to use not the statistics of the variants itself but the underlying rules that were applied during the segmentation process of MAUS. Note that this has nothing to do with the statistical weights of the microrules mentioned earlier in this paper; it's the number of appliances of these rules that counts. Since there is formally no distinction between microrules for segmentation in MAUS and probabilistic rules for recognition, we can use the same format and formalism for this approach as in MAUS. The step-by-step procedure is as follows:

A: Derive a set of statistical microrules from a subset of manually segmented data or use the rule set PHONRUL (see section 1).

B: Apply this rule set to segment the training corpus and count all appliances of each rule forming the statistics of the recognition rule set.

Note that the recognition rule set might be a subset of the PHONRUL/microrule set, although this is very unlikely for the latter.

This approach has the great advantage that the statistics are more compact (and therefore robust), independent of the dictionary used for recognition (which for sure will contain words that were never seen in the training set) and generalise knowledge about pronunciation to unseen cases. However, the last point may be a source of uncertainty, since it cannot be foreseen whether the generalisation is valid to all cases where the context matches. We cannot be sure that the context we are using is sufficient to justify the usage of a certain rule in all places where this context occurs.

4. AUTOMATIC SPEECH RECOGNITION (ASR)

There have been several attempts to incorporate knowledge about pronunciation into standard methods for ASR. Most of them (with some exceptions, e.g. [4]) didn't yield any improvements. The argument was that the advantage of a better modelling on the lexical level is eaten up by the fact that the search space and/or the dictionary ambivalence increases. However, most of the literature did not take into account reliable statistics (because they were simply not available) and used acoustic models that were trained using canonical pronunciations. Our hypothesis is that an increase in recognition performance can only be achieved if the following two conditions are satisfied:

1. A reliable statistical model for pronunciation (which very likely will be adapted to the task).
2. Acoustical models that match the modelling on the lexical level.

On this basis we are currently conducting several experiments with a standard HTK recogniser for the 1996 Verbmobil evaluation task. In this paper we will only report about preliminary results using the direct statistics approach of section 3.1.

A standard recogniser of HTK 2.0 with the following properties was designed for the experiment:

The speech signal is mean subtracted, emphasised and preprocessed into 12 MFCCs + log Energy, Delta, Delta-delta every 10 msec. Training and test sets are defined in the 1996 Verbmobil evaluation task ('Kuer', test corpus: 6555 words). The canonical dictionary contains 840 different entries. The language model is a simple bigram calculated exclusively from the training set. The acoustic models are monophone left-to-right HMMs with 3-5 states containing a variable number of mixtures without tying. We use 46 models from the extended German SAM-PA including one model for silence and one model for non-speech events.

We trained and tested the recogniser with the same amount of data in two different fashions:

- *Baseline System*
  Standard bootstrapping to manually labelled data (1h40) and iterative embedded re-estimation (segmental-k-means) using 30h of speech until the performance on the independent test set converged (note: performance in terms of word accuracy, defined by (number of words - insertions - replacements - deletions) / number of words). The re-estimation process used a canonical pronunciation dictionary with one pronunciation per lexical entry.
  The system was tested with the same canonical dictionary.

- *MAUS System*
  This system was bootstrapped to one third of the training corpus (approx. 10h of speech) using the MAUS segmentation and then iteratively re-estimated (30h of speech) using not the canonical dictionary but the transcripts of the MAUS analysis (note that the segmental information of the MAUS analysis is NOT used for the re-estimation).

The system was tested with the probabilistic pronunciation model described in section III.1. using the pruning parameters N=20 and M=0%.

Figure 3 shows the performance of both systems during the training process. Note that the MAUS system starts with a much higher performance because it was bootstrapped to 10h of MAUS data (compared to 1h40min of manually labelled data for the baseline system). After training, the MAUS system converges on a significantly higher performance level of 66.35% compared to 63.44% of the baseline system.

5. CONCLUSION

The MAUS system can be used effectively to fully automatically label and segment read and spontaneous speech corpora into broad phonetic alphabets. This enables us for the first time to derive statistical models on different processing levels (acoustic, phonetic, lexical) on the basis of very large databases. We have shown that the usage of this data can significantly improve ASR on spontaneous speech.

The MAUS principle is not language dependent (however, the required resources are!). Therefore we strongly encourage colleagues in other European countries to adopt the MAUS principle for their specific languages and produce similar resources as are currently produced at BAS for the German language. A first joint project (MIGHTY MAUS) for American English and Japanese is scheduled for 1998 together with the International Computer Science Institute (ICSI), Berkeley California, and Sofia University, Tokyo.

REFERENCES

[1] A. Kipp. M.-B. Wesenick, F. Schiel (1997): Pronunciations Modelling Applied to Automatic Segmentation of Spontaneous Speech; in Proceedings of the EUROSPEECH 1997 Rhodes, Greece, pp. 1023-1026.

[2] A. Kipp, M.-B. Wesenick, F. Schiel (1996): Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora; in: Proceedings of the ICSLP 1996. Philadelphia, pp. 106-109.

[3] A. Kipp, M.-B. Wesenick (1996): Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals; in: Proceedings of the ICSLP 1996. Philadelphia, pp. 129-132.

[4] T. Sloboda, A. Waibel (1996): Dictionary Learning for Spontaneous Speech; in: Proceedings of the ICSLP Philadelphia, pp. 2328.

[5] SAM Phonetic Alphabet, www.phon.ucl.ac.uk/home/sampa/home.htm

[6] IPDS (ed.): The Kiel Corpus of Spontaneous Speech; CDROM 1 + 2, University of Kiel, 1995.

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

[8] PHONOLEX, www.phonetik.uni-muenchen.de/Bas/BasPHONOLEXeng.html or
F. Schiel (1997): The Bavarian Archive for Speech Signals; in: FIPKM 1997, Institute of Phonetics, University of Munich, to appear.

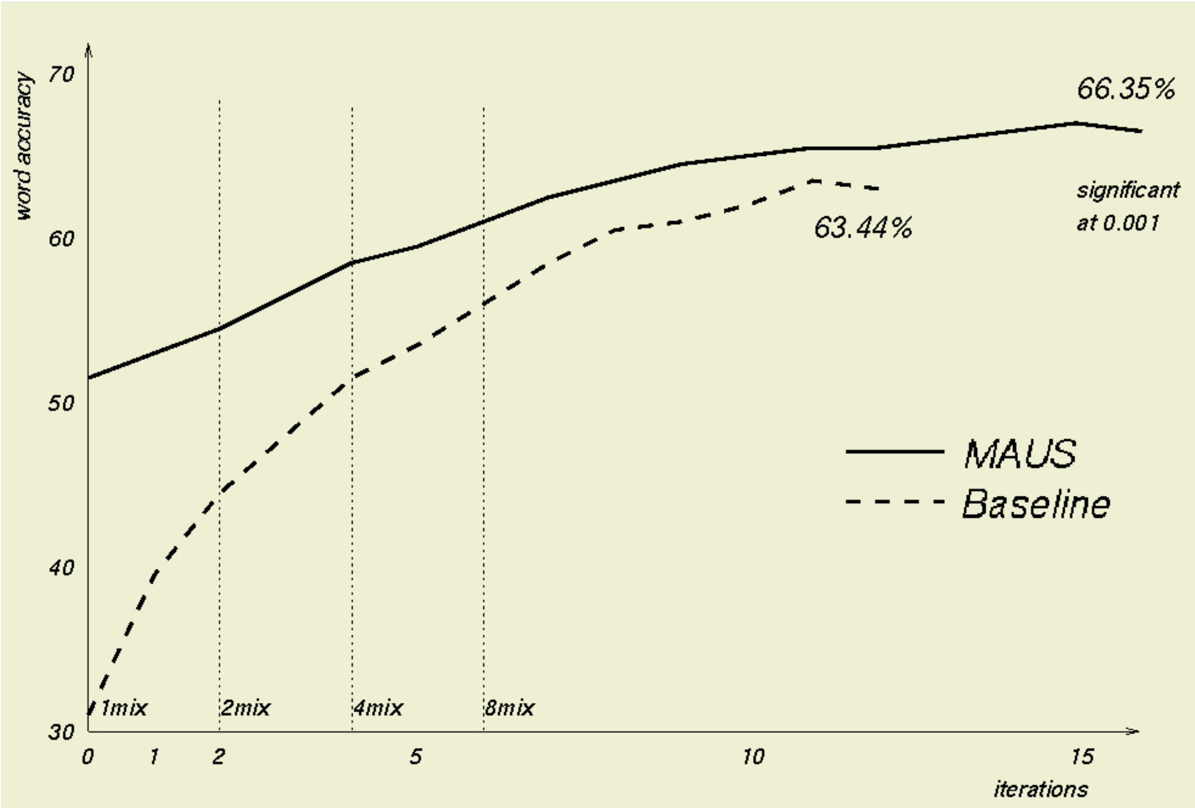

Figure 3: Performance of baseline system compared to the system trained with MAUS data and probabilistic pronunciation model

[9] S. Young et al. (1995/1996): The HTK Book; Cambridge University.

[10] Partitur Format, www.phonetik.uni-muenchen.de/Bas/BasFormatsdeu.html
or
F. Schiel, S. Burger, A. Geumann, K. Weilhammer (1997): The Partitur Format at BAS. In: FIPKM 97, Institute of Phonetics, University of Munich, to appear.