Phonemic Segmentation and Labelling using the MAUS Technique

Florian Schiel, Christoph Draxler, Jonathan Harrington

Bavarian Archive for Speech Signals, Institute for Phonetics and Speech Processing,
Ludwig-Maximilians-Universität, München, Germany
schiel|draxler|jmh@bas.uni-muenchen.de

Abstract

We describe the pronunciation model of the automatic segmentation technique *MAUS* based on a data-driven Markov process and a new evaluation measure for phonemic transcripts *relative symmetric accuracy*; results are given for the MAUS segmentation and labelling on German dialog speech. MAUS is currently distributed as a freeware package by the Bavarian Archive for Speech Signals and will also be implemented as a web-service in the near future.

Index Terms: MAUS, phonemic segmentation, automatic segmentation, pronunciation model, data-driven, relative symmetric accuracy, freeware, web-service

1. Introduction

Phonemic segmentation and labelling (S&L) of speech corpora is required for a number of phonetic analysis and technical tasks. Manual segmentations are precise but inconsistent, since they are often produced by more than one labeler, and require time and money. Automatic S&L systems generate reproducible results, are much faster (often realtime), but not as precise as human labelers. Nevertheless project requirements often dictate the usage of automatic methods. Practical applications of automatic S&L are nowadays always implemented as a statistical search for a S&L $\hat{K}$ in a space Ψ of all possible S&Ls, which can be formulated as:

$$\hat{K} = \text{argmax}_{K \in \Psi} P(K|O) = \text{argmax}_{K \in \Psi} \frac{P(K)p(O|K)}{P(O)}$$

where O is the acoustic observation on the corresponding speech signal. Since the probability for the observation $P(O)$ is a constant for all K this can be reduced to the simple well known formula

$$\hat{K} = \text{argmax}_{K \in \Psi} P(K)p(O|K)$$

where $p(O|K)$ models the probability (density) of the acoustics given a certain (discrete) S&L (e.g. by using HMM, ANN etc.) while $P(K)$ models the probability of the symbol sequence in the S&L K ([1]).

Automatic S&L systems mainly differ in Ψ and the way that $P(K)$ is modeled. For example a simple *forced alignment* to a given phonemic transcript yields

$$||\Psi|| = 1 \quad \text{and} \quad P(K) = 1$$

and hence only $p(O|K)$ is maximized here.

$P(K)$ does not necessarily need to be a statistical model. For instance in [2] and [3] Ψ was determined by applying phonological pronunciation rules to a canonical pronunciation form yielding M pronunciation variants which were then treated with the same probability $P(K) = \frac{1}{M}$. Other ways to model $P(K)$ are the usage of an n-gram phonotactic model, a lexicon of pronunciation variants or a Markov process, which is the MAUS method.

The MAUS system models $P(K)$ for each recording O by building an acyclic directed graph $\mathcal{G}(N, A)$ with phonemic symbols in the nodes N and transition probabilities on the arcs A. Each path from the start node to the end node represents a possible $K \in \Psi$ and accumulates to the probability $P(K)$. $p(O|K)$ is determined by HMMs for each phonemic segment and a simple Viterbi search through the graph yields the maximal $P(K)p(O|K)$ and by backtracking the path through $\mathcal{G}$ $\hat{K}$ is determined ([4]).

In this presentation we will concentrate on the technique to build the core pronunciation model used by MAUS and how we extract data-driven, statistically weighted pronunciation rules from an annotated speech corpus for that purpose (Section 2). Section 3 will discuss the evaluation of S&L systems and give some results from MAUS, while Section 4 describes the tools of the MAUS freeware package and the supported languages.

2. Pronunciation Model

2.1. Building the Automaton

Input to the process is a string of orthographic words representing the spoken utterance[1]. The orthographic form is transformed into a citation pronunciation form, called the *canonical form* $\mathcal{C}$ hereafter. This can be done either by lexicon lookup or a text-to-phoneme system, or - as in the case of MAUS - a combination of both. The canonical form $\mathcal{C}$ can be represented by a simple left-to-right finite-state automaton $\mathcal{G}_c(N, A)$ without self transitions where each node emits exactly one phonemic symbol; the first and last states are non-emitting enter and exit states.

$\mathcal{G}_c$ can now be extended by additional arcs, emitting and non-emitting states to model variations from the canonical form. Technically this is done by applying a set of matching substitution rules where each rule is defined by a tuple (a, b, l, r) with a pattern string a, a replacement string b and left/right context strings l, r. Essentially each application of a rule creates a new arc with a number of new nodes (or zero). a, b, r, l may also be the empty symbol $\emptyset$ to allow for insertions and deletions of symbols as well as non-defined contexts. In addition the symbol /#/ may be used to model word boundaries, to allow the modeling of cross-word effects or word initial/final contexts. Since substitution rules are only applicable to the canonical form (the sub-automaton $\mathcal{G}_c$) a single pass over the rule set creates an automaton $\mathcal{G}$ covering all possible pronunciation variants (with no recursive applications of rules required).

[1] E.g. taken from an orthographic transcription of the speech corpus.

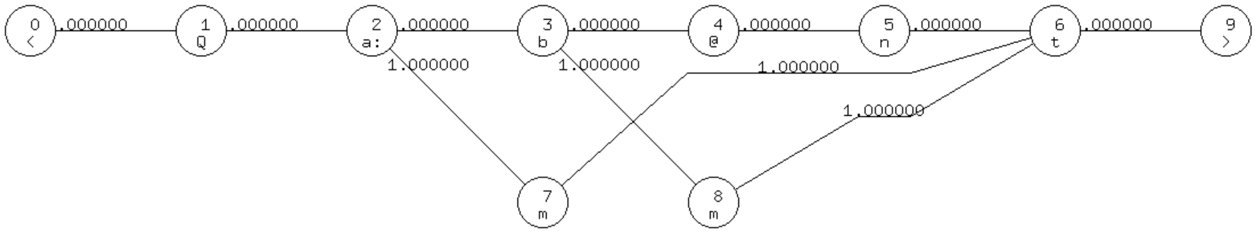

Figure 1: *Example automaton for the word 'Abend'. /</ and />/ are non-emitting states.*

Consider for instance the canonical form of the German word 'Abend'[2]:

/? a: b @ n t/

To model the very common reduction/assimilation processes that lead to the realizations

/? a: b m t/ and /? a: m t/

we need two substitution rules:

(/@n/,/m/,/b/,/t/) (/b@n/,/m/,/a:/,/t/)

resulting in the automaton as shown in Figure 1.

2.2. Creating the Rule Set

In principle the set of substitution rules can be derived from different sources: they can be either data-driven (see below) or hand-crafted. In the latter case the rule set often represents a phonological model of the language concerned and contains no information about the probabilities of its possible substitutions. Although this most likely deteriorates the performance of the S&L we sometimes use this technique for languages where there is not enough annotated training material available or when dealing with special speech recordings documenting well-known phonological processes as is quite common in phonetic studies.

MAUS achieves the best performance when using a data-driven statistical weighted rule set. Rules (a, b, l, r) can be found by performing a *longest common subsequent alignment* ([1]) between the canonical form $\mathcal{C}$ and the annotation (the realization) $\mathcal{R}$ of a recorded utterance and then segmenting the alignment for common and deviating portions. When restricting the left/right context l, r of each rule to length 1, it is quite straightforward to extract rules (a, b, l, r) from each deviating portion of the alignment and determine their total number $n(a, b, l, r)$ from the annotated corpus. In parallel the total number of occurrences of the string (l, a, r) can by derived from the canonical forms $\mathcal{C}$ of the corpus: $n(l, a, r)$. Using maximum-likelihood we can then estimate the conditional probability of the application of a rule by:

$$\hat{P}(b|l, a, r) = \frac{n(a, b, l, r)}{n(l, a, r)}$$

Since annotated speech corpora are rare and in most cases small, simple maximum-likelihood estimates may not generalize sufficiently. There are two possible ways to yield a more robust rule set:

1. Use a discounting technique to spread probability mass to all unseen rule contexts (l, a, r). This leads to an explosion of the rule set and subsequent computational problems. Therefore it is necessary to restrict the discounting by pruning to an under-specified set of phonological rules.
2. Split each rule with non-empty left and right context into two left/right independent rules and discount probability mass to these unseen but plausible new rules. The basic idea here is that since left and right context might be statistically independent, the system might encounter pronunciation variants with only either the left or the right context or new combinations of those.

In the current MAUS system we use the second approach, since it proved to be more robust than the brute force discounting technique.

2.3. From Automaton to Markov Model

Up to this point we have created a finite-state automaton that covers all hypothetical realizations predicted by the rule set. To use this automaton effectively for a combined acoustical/phonotactic Viterbi search, we need to augment it by probabilities for emissions and transitions, thus creating a true Markov process. This is not a trivial task since the automaton may model paths (= phoneme sequences) of different length, but still every path K through the model must yield to the appropriate accumulated probability $P(K)$.

Without loss of generality we can define the *emission probability* of each node as the production probability of a Hidden Markov Model (HMM) that is trained to manually segmented training samples of the corresponding label class. In other words we replace the nodes N of $\mathcal{G}(N, A)$ by class-corresponding HMMs.

Regarding the *transition probabilities* between nodes we have to distinguish two cases:

1. Accumulated probabilities are equally distributed over all hypothesized realizations $\mathcal{R} \in \Psi$.
   In this case no statistical information about rule application is available.
2. Accumulated probabilities must reflect the $\hat{P}(b|l, a, r)$ of all applied substitution rules (a, b, l, r) along a path $\mathcal{R}$ through the model.
   In this case we use data-driven statistically weighted rules as described above.

Due to the space constraints of a short paper we will only demonstrate the first case and will also not consider the additional problem of overlapping contexts from different rules; the second case can be done in analogy (see [1] for details).

We define the *rank* of a node d_i as its distance from the non-emitting starting node (the starting node has rank 0), the set $\Gamma^{-}(d_i)$ as the set of all nodes that precede a node d_i and

[2] Phonemic symbols in SAM-PA.

$\Gamma^+(d_i)$ as the set of nodes that follow d_i. Let $N(d_i)$ be the number of possible paths that end in node d_i, which equals the sum of all paths ending in preceding nodes of d_j. $N(d_i)$ can therefore be calculated for all nodes in ascending rank order by applying the recursive formula

$$N(d_j) = \begin{cases} 1 & \text{for the starting node} \\ \sum_{d_i \in \Gamma^-(d_j)} N(d_i) & \text{else} \end{cases}$$

$P(d_i)$, the probability that a node is part of a phoneme sequence can also be calculated for all nodes, since we know that this probability must be 1 for the last node, and we can recursively calculate the probabilities with descending rank order using

$$P(d_i) = \sum_{d_j \in \Gamma^+(d_i)} P(d_j) P(d_i|d_j) = \sum_{d_j \in \Gamma^+(d_i)} P(d_j) \frac{N(d_i)}{N(d_j)}$$

Since the model is acyclic and we consider all paths through the model as equally probable, we can say that the backward probability that a node d_i precedes a node d_j is

$$P(d_i|d_j) = \frac{N(d_i)}{N(d_j)} \quad \text{with} \quad d_i \in \Gamma^-(d_j)$$

By applying Bayes we get the desired transition probability from node d_i to node d_j:

$$P(d_j|d_i) = \frac{P(d_j)N(d_i)}{P(d_i)N(d_j)} \quad \text{with} \quad d_i \in \Gamma^-(d_j)$$

which then can be calculated for each transition found in the model.

3. Evaluation of S&L Systems

Evaluating the quality of S&L systems is usually done separately for the label sequence and the segmental boundaries.

3.1. Evaluation of the Label Sequence

In the literature Cohen's κ statistics[3] is often used for inter-labeler agreement or the evaluation of S&L systems ([8]). κ is a quality measure for how much a labelling overlaps with regard to a gold standard while at the same time being independent of the size of the label inventory. We do not consider κ an appropriate measure for automatic S&L because it does not reflect the reality of S&L: there is no gold standard for phonemic S&L (even the best phoneticians disagree about a considerable proportion of cases), usually a fixed-sized phoneme inventory is applied (and hence independence of the inventory is not required), and κ does not allow for the inter-labeler agreement of human transcribers on the same task (task difficulty).

Instead we propose the *relative symmetric accuracy* (RSA), which calculates the ratio of mean symmetric system-to-labeler agreement to mean symmetric inter-labeler agreement.

Let $A(\mathcal{S}_1, \mathcal{S}_2)$ be the accuracy between two symbol sequences $\mathcal{S}_1$ and $\mathcal{S}_2$ measured from the alignment errors derived from a *longest common subsequent alignment*:

$$A(\mathcal{S}_1, \mathcal{S}_2) = \frac{N_1 - N_{del} - N_{ins} - N_{rep}}{N_1}$$

with

N_1	:	number of symbols in $\mathcal{S}_1$
N_{del}	:	number of deletions
N_{ins}	:	number of insertions
N_{rep}	:	number of replacements

Since $A(\mathcal{S}_1, \mathcal{S}_2) \neq A(\mathcal{S}_2, \mathcal{S}_1)$, the *symmetric accuracy* is defined as ([1]):

$$SA = \frac{A(\mathcal{S}_1, \mathcal{S}_2) + A(\mathcal{S}_2, \mathcal{S}_1)}{2}$$

Let's assume that we have S&Ls from a group of human labelers on the same test set material and one S&L from the system to be evaluated. Then we can calculate the mean symmetric accuracy $\widehat{SA}_{hh}$ over all possible pairings between human labelers and the mean symmetric accuracy of all possible pairings between human labelers and the system $\widehat{SA}_{hs}$. The *relative symmetric accuracy* is then simply:

$$RSA = \frac{\widehat{SA}_{hs}}{\widehat{SA}_{hh}} 100\%$$

Note that RSA may be greater than 100%, if the system outperforms the human labeler group[4].

The MAUS system[5] has been evaluated using the described technique on a manually segmented and labelled sub-portion of the German Verbmobil 1 corpus[6]. Three independent human labelers produced S&Ls for two complete dialogs with a total of 9587 phonemic symbols in the canonical form. The mean symmetric accuracy between all human labeler pairings was $\widehat{SA}_{hh} = 84.01\%$; the mean symmetric accuracy between human labelers and system was $\widehat{SA}_{hs} = 81.85\%$. The RSA for German MAUS on spontaneous speech is therefore:

$$RSA = 97.43\%$$

3.2. Evaluation of Segmentation

Figure 2 shows a MAUS segmented signal taken randomly from the German ALC corpus of intoxicated speakers ([11]).

There exists no widely accepted methodology to evaluate the quality of a phonemic segmentation with regard to a reference segmentation. The same problem of a missing gold standard as discussed above is also an issue here – and to make things worse, phonemic symbol sequences may differ between segmentations and hence a simple mapping of corresponding segmental boundaries in both segmentations is not feasible.

Often only segmental boundaries of corresponding phonemic segments in both S&L are taken into account and counted for deviations above a threshold, e.g. 20msec. A better alternative is a histogram over the deviation as shown in Figure 3 for the German MAUS system. Note that the center of the Gaussian-like distribution is not at zero; this effect has been observed for many HMM-based segmentation systems. The shift is usually about the size of one window in the front-end processing of the recognizer. The reason for this shift is still unclear; in MAUS we simply compensate for this shift with a counter-shift of 10msec.

In general, automatic segmentations lack the accuracy of a trained phonetician. Studies dealing with durations of linguistic or sub-linguistic events (e.g. voice onset time) require a manual correction step before exploiting the results. However, automatic segmentations may be successfully applied to locate linguistic entities such as phones, syllables, morphs or words, for instance to measure fundamental frequency, formants, spectral shapes etc.

[3] or Fleiss' κ for more than one labeller ([9])

[4] This is also a feature that the κ statistics does not provide.

[5] The German MAUS system has been trained on the *Kiel Corpus*.

[6] Spontaneous speech, dialogues *m116d* and *m231d*, see [10]

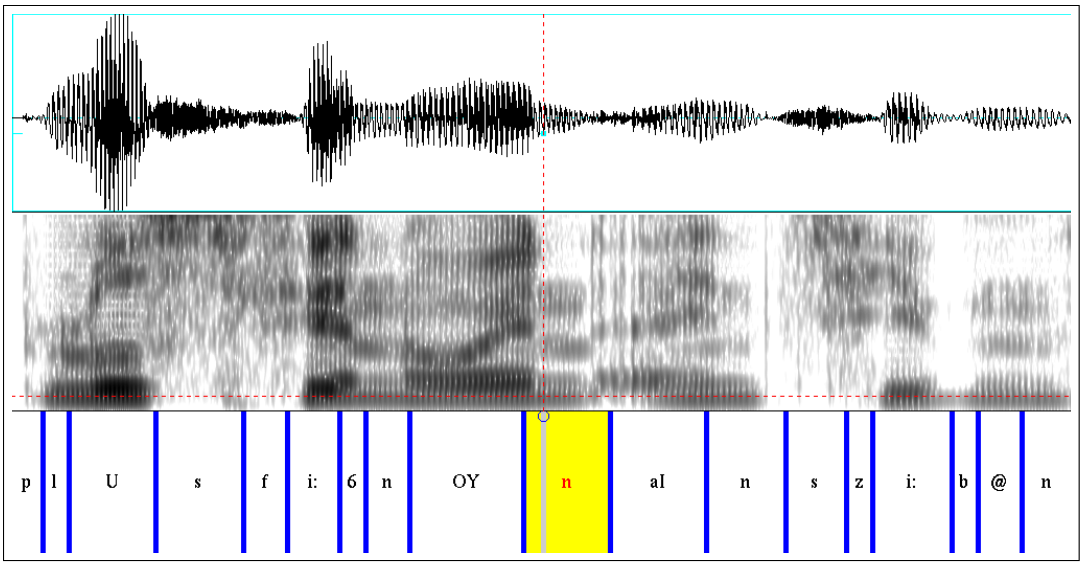

Figure 2: *Example MAUS segmentation and labelling taken from the German ALC corpus (phonemic labels in SAM-PA).*

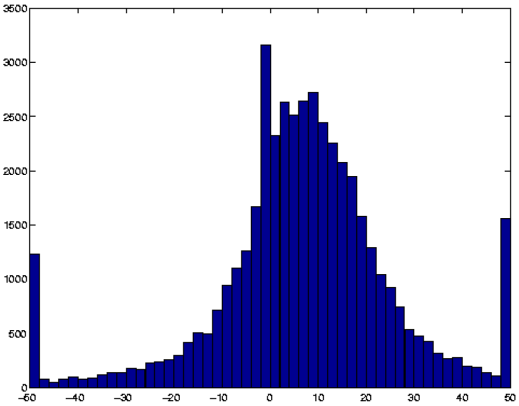

Figure 3: *Histogram of boundary deviations in msec of a German MAUS S&L evaluation*

4. MAUS Software Package

The MAUS software package is available as freeware at the Bavarian Archive for Speech Signals ([5]). Version 2.17 covers the languages German, English, Italian, Estonian, Hungarian, Spanish (beta) and Portuguese (beta).

The main tool *maus* performs an automatic S&L on a single recording, starting with the orthographic transcript. Aside from the standard output format in BAS Partitur Format (BPF)[7] *maus* also provides Emu[8] or praat[9] compatible output. Beside the basic tool the MAUS package comprises the tools *maus.corpus* for the S&L of whole corpora and *maus.iter* for adapting the acoustical model to the input data. MAUS can be adapted to new languages by mapping the phonemic symbol set, formulating a new pronunciation rule set and adapting the HMM using *maus.iter* to a training corpus.

The system requirements for running MAUS are System V UNIX OS, csh and awk interpreter, gcc and HTK[10]. To simplify the usage of MAUS we implemented a web-service where prospective users can upload signal files and receive the S&L in return. This web-service is now in alpha and will be released within the CLARIN project ([7]) by the end of 2011, together with a number of other speech processing tools within the WikiSpeech infrastructure ([6]).

5. References

[1] Kipp A (1998): Automatische Segmentierung und Etikettierung von Spontansprache. *Doctoral Thesis*, Technical University Munich.

[2] Wester M, Kessens J M, Strik H (1998): Improving the performance of a Dutch CSR by modeling pronunciation variation. *Workshop on Modeling Pronunciation Variation*, Rolduc, Netherlands, pp. 145-150.

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

[4] Schiel F (1999) Automatic Phonetic Transcription of Non-Prompted Speech. *Proceedings of the ICPhS*, San Francisco, August 1999. pp. 607-610.

[5] MAUS: *ftp://ftp.bas.uni-muenchen.de/pub/BAS/SOFTW/MAUS*

[6] Draxler Chr, Jänsch K (2008): WikiSpeech – A Content Management System for Speech Databases. *Proceedings of Interspeech*, Brisbane, Australia, pp. 1646-1649.

[7] CLARIN: *http://www.clarin.eu/*

[8] Cohen J (1960): A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1): 37-46.

[9] Fleiss J L (1971): Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, No. 5 pp. 378-382.

[10] Burger S, Weilhammer K, Schiel F, Tillmann H G (2000): Verbmobil Data Collection and Annotation. In: *Verbmobil: Foundations of Speech-to-Speech Translation* (Ed. Wahlster W), Springer, Berlin, Heidelberg.

[11] Schiel F, Heinrich Chr, Barfüßer S (2011): Alcohol Language Corpus. *Language Resources and Evaluation*, Springer, Berlin, New York, in print.

[7] *http://www.bas.uni-muenchen.de/Bas/BasFormatseng.html*

[8] *http://emu.sourceforge.net/*

[9] *http://www.praat.org/*

[10] *http://htk.eng.cam.ac.uk/*