University of Szeged
Research Group on Artificial Intelligence

Posterior-Based Speech Models and their Application to Hungarian Speech Recognition

Summary of the PhD Dissertation

by

László Tóth

Advisor:

Prof. Dr. János Csirik

Szeged
2006

Introduction

The current speech recognition technology is built on statistical principles instead of speech-specific knowledge. Although there are constant attempts to incorporate what is known about human speech perception, these usually refine only the preprocessing step and leave the statistical framework untouched. In particular, the 3-state left-to-right hidden Markov phone modelling (HMM) methodology has been practically unchallenged for the last decade. Rather, development efforts have focused mainly on collecting enormous training corpora and on building sophisticated language models. However, nowadays the technology seems to have reached its limits, its abilities still being far from that of humans. We think that it is time to step back and refine the acoustic models as well, retaining the statistical approach but narrowing the gap between the properties of the models and human hearing.

This dissertation starts out by gathering critical remarks on hidden Markov models from a speech perception point of view, and we also discuss some general properties of an envisioned alternative recognition framework. After, we point out that the main issue of statistical modelling in speech recognition is that the utterance-level probabilities have to be decomposed into the probabilities of some smaller units. Unfortunately, probability theory offers only a very limited range of tools for decomposing a multivariate probability. Yet, by applying these in different ways, we can construct various kinds of speech models that differ from HMMs in several aspects. A common property of the models we will use here is that they combine posterior probabilities, while the HMM builds its scores from class-conditional likelihoods. As posterior probability estimators we apply neural networks (ANN) – this approach having a number of advantages compared to modelling class-conditional likelihoods by Gaussian mixtures, as is usual in HMMs.

The most dubious feature of the hidden Markov model is that its probability decomposition goes down to the level of speech frames – which are then assumed to be conditionally independent, and their likelihood values are combined by multiplication. We propose two alternative decompositions that avoid this so-called 'naive Bayes' assumption. In one case we simply do not decompose the phonetic segments into frames, but instead model them as one unit. This approach leads to the family of segment-based models, and a significant part of the dissertation deals with the issues of how to parametrize and train these. As it turns out, they have some distinct advantages – for example, they are much better at classifying phonetic segments – but their particular drawback is that they have difficulties in finding the proper segmentation during decoding. Hence, we will suggest and test various methods to overcome this problem.

Seeing the special problems of the segment-based framework, in the other model examined we return to the conventional frame-based approach, but now we try to combine the frame-based posterior probabilities via averaging instead of multiplication. Although this may sound nonsensical at first, we introduce several arguments for it from classifier combination literature. The experiments show that in classification tasks averaging is indeed no worse than multiplication. However, to enable it to perform phonetic decoding, it has to be extended with a segmentation probability component – a lesson learned from the segmental model approach. We will call the resulting framework the 'averaging HMM/ANN hybrid'. In all the experiments we do – phone recognition and isolated word recognition tasks – the averaging model performs the same as or better than the conventional HMM. In a subsequent chapter we improve its performance even more by extending it with an explicit duration model and a resampling-based training scheme.

A notable feature of the dissertation is its segment-based view on the decoding task. It originates from the experiments with the segment-based model, but we apply it to the frame-based systems as well. Most importantly, it turns our attention to the question of how the frame-based models solve the segmentation problem of the decoding task. Both the conventional HMM and the HMM/ANN hybrid will be examined from a segment-based point of view, and in both cases we will conclude that it is basically the multiplication-based combination of the frames that enables these models to hypothesize reasonable segmentations. This insight gained into the working of the frame-based systems is probably a more important result of this dissertation than the two posterior-based models suggested and studied.

All the recognition experiments of this dissertation will be carried out on Hungarian speech databases. As in most cases there will be no comparative results available, the hidden Markov model toolkit (HTK) will be used to obtain a basis for comparison.

The Decomposition Problem in Statistical Speech Recognition

Statistical pattern recognition is the most successful approach to machine learning. According to its Bayes decision rule, for optimal classification we have to know the posterior probability $P(W|X)$ for every possible class label W and observation vector X. There are many techniques available to obtain an estimate of $P(W|X)$, and these have been successfully applied to many practical problems. Speech recognition, however, is special in the sense that the number of possible observations (say, speech signals said in one breath) and class labels (phone or word series) is too huge to be managed directly in one step. Rather, it is necessary to decompose both the observation vector and the class label into some smaller units. Direct probability estimates are then created only over the subspaces of these units, and an estimate of the global probability is created by properly combining the local estimates of the units. Thus, crucial issues are how we select the local recognition units, what technology we use to model their probabilities, and how we combine their probability estimates into a global estimate.

Probability theory offers only a few techniques for manipulating and decomposing a multivariate probability in a simple way. One of them is the Bayes rule that changes the role of the variables in a conditional probability. Thanks to it, we can choose between modelling $P(W|X)$ or $p(X|W)$. As the first one is a discrete probability while the latter one is a continuous probability density, modelling them requires different technologies. The class posteriors $P(W|X)$ are usually estimated by neural networks, while the class-conditional densities $p(X|W)$ are estimated by Gaussian mixtures. In speech recognition the latter – so-called generative – approach is currently preferred.

The second important decomposition technique is to introduce a latent variable, the values of which form a set of disjoint events. In speech recognition we assume that the signal can be regarded as a series of phonetic segments, and we use the phonetic segmentation S as a latent variable. Then, according to the law of total probability,

$$p(X|W) = \sum_S p(X,S|W) = \sum_S p(X|S,W)P(S|W) \approx \max_S p(X|S,W)P(S|W), \quad (1)$$

where in the second step the chain rule was applied. In practice the summation is usually approximated by a maximization, which turns the recognition task into a search problem over all possible W and S.

The third decomposition trick we use is the assumption of independence (also known as the naive Bayes assumption), which may can be applied at two levels. First, we decompose the utterance into acoustic-phonetic data segments X_i, segment boundary pairs $S_i = (s_{i-1}, s_i)$ and phone labels u_i:

$$p(X|S,W) \approx \prod_i p(X_i|u_i) \quad \text{and} \quad P(S|W) \approx \prod_i P(S_i|u_i). \tag{2}$$

In conventional HMM speech recognizers the segments are decomposed one step further. The acoustic data X_i of the segment is represented as a series of data frames $(x_{s_{i-1}}, ..., x_{s_i - 1})$, and the corresponding likelihood is estimated as

$$p(X_i|u_i) \approx \prod_{j=s_{i-1}}^{s_i - 1} p(x_j|u_i), \tag{3}$$

while $P(S_i|u_i)$ is approximated by a geometric duration model (defined by the state transition probabilities).

Although the above derivation is mathematically attractive, almost all of its steps can be criticized from a speech perception point of view. The last decomposition step (Eq. (3)) is the most debatable. For example, it is now known that humans process speech in units that are longer in time and narrower in frequency than the conventional data frames [9]. The conditional independence assumption of the frames can also be argued against from several aspects [13]. Finally, working with class-conditionals as Eq. (3) does, is not necessarily the best solution from a technical point of view. A scheme that combines posteriors would allow the use of ANN estimates, and in practice these were found to have a better discrimination ability and greater flexibility [4].

In the following we examine two decompositions that both result in posterior-based models. In one of them the decomposition into frames will be avoided by modelling whole segments. The other model will work with frames, but combines their posterior estimates by averaging instead of multiplication.

A Posterior-based Segmental Speech Recognition Model

The main motivation for creating segment-based models is to eliminate the false conditional independence assumption (Eq. (3)) of the frames. Most of the proposed segmental descriptions are generative [13], and the posterior-based solutions are relatively rare [1; 12; 21]. The key difference between the two approaches is that while in the former the decomposition takes the form of Eq. (1), in the latter we have

$$P(W|X) = \sum_S P(W,S|X) = \sum_S P(W|S,X)P(S|X) \approx \max_S P(W|S,X)P(S|X). \tag{4}$$

The resulting components are then decomposed into segment-level values like so

$$P(W|X) \approx \prod_i P(u_i|X_i) \quad \text{and} \quad P(S|X) \approx \prod_i P(S_i|X_i). \tag{5}$$

The task of $P(u_i|X_i)$ is to associate phone probabilities with the segments, and hence we will call it the *phone classifier*. The *segment probability* $P(S_i|X_i)$ has to tell us how likely the given signal

Classification error rate	
Baseline features	Baseline plus duration
47.72%	42.15%

Off-line normalization methods	CER%
Mean and dev. normalization (full spectrum)	40.27%
Mean and dev. normalization (per channel)	37.75%
On-line normalization methods	
RASTA filtering	43.86%
Mean and dev. norm. (per channel, $\tau = 250ms$)	41.12%
Mean and dev. norm. (per channel, $\tau = 1sec$)	40.36%
Nonlinear AGC (per channel, $\tau = 250ms$)	39.64%
Nonlinear AGC (per channel, $\tau = 1sec$)	38.49%

	CER with observation context of τ		
Normalization	$\tau = 150msec$	$\tau = 250msec$	$\tau = 1sec$
Off-line mean and dev. norm.	33.18%	34.49%	36.12%
Nonlinear AGC (*1sec*)	33.51%	34.85%	36.25%

CER with onset/offset features (off-line norm., 150ms obs.cont.)	32.17%

Table 1: Phone classification error rates on the MTBA corpus

segment corresponds to a phonetic segment. The $P(S|X)$ value formed from these will be referred to as the *segmentation probability*.

If we intend to implement the phone classifier by means of neural networks, then we have to describe every segment with the same number of features, independent of its duration. The simplest way of creating such a feature set is to take conventional frame-based features and represent the segment with the averages of these over the segment. We applied this technique to the energy of the signal calculated in Bark frequency bands, and averaging was performed over each third of the segment. This resulted in our baseline feature set. It was extended with several additional features such as the duration of the segment and the average band energies calculated over the observation context of the segment. We also experimented with onset and offset detector features that measure the degree of change at the segment boundaries. To normalize the signal, both adaptive gain control (AGC) algorithms and a normalization of the mean and variance of the energy trajectories were tested. The feature parameters were fine-tuned on the MTBA Hungarian Telephone Speech Database that contains phonetically rich sentences recorded via telephone lines. Table 1 shows how the step-by-step introduction of the features improved the phone classification error rate on this corpus. Although examples of a similar set of energy features can be found in the literature [6], most of the additional features that we introduced are quite distinct.

A similar feature set was applied in the phone classification experiments on the OASIS-Numbers database. This database contains recordings of numbers only. In this case we have comparative results from an HMM and from an SVM (the latter also using the segmental features). The scores clearly reflect the superiority of the posterior-based models (ANN and SVM) over the HMM (see Table 2).

HMM	ANN	SVM
9.34%	7.78%	5.81%

Table 2: Phone classification error rates on the OASIS-Numbers corpus

Obtaining a good estimate for the segmentation probability component $P(S|X)$ is much more difficult than classifying the segments. We can find examples in the literature where the estimation of this component is simply avoided by running an HMM recognizer to obtain the N most probable segmentations [1]. Another option is to estimate, for each frame, the probability of them being a segment boundary position. These frame-based scores can then combined to obtain an estimate of $P(S|X)$ [12]. For the sake of computational efficiency, we looked for a solution that makes our segmental phone classifier ANNs capable of handling $P(S_i|X_i)$ as well.

To understand how $P(S_i|X_i)$ can be interpreted from a segment-based view, consider the fact that during recognition the decoder encounters segments that do no correspond to real phones. The phone classifier is not automatically able to detect and report these segments, since it is neither trained to do so nor has an output for them. These segments are outliers from the phone classification point of view, or – borrowing the terminology of Glass et al. – they are 'anti-phones' [6]. To enable the phone classifier neural net to handle them, it has to be extended with an additional class for these segments, and examples should also be generated for this class in the training phase. To create such training examples we took the real phonetic segments of a manually segmented corpus and shifted their boundaries in both directions by 30-30 ms. With the help of the shifted and the real boundaries six anti-phone examples were created from each phone example during the training process.

Having obtained segment-level estimates $P(S_i|X_i)$, we can get an estimate of $P(S|X)$ by using

$$P(S|X) \approx \prod_i P(S_i|X_i). \tag{6}$$

Unfortunately, in practice we found that this formula does not guarantee a proper normalization between segmentations. Glass et al. suggest that better results can be obtained if the formulation includes not just the segments of S, but all other segments that the recognizer encounters during the decoding process. Based on this concept, we arrive at the approximation

$$P(S|X) \approx \prod_i P(S_i|X_i) \prod_{s \in \overline{S}} (1 - P(s|X(s))), \tag{7}$$

where $\overline{S}$ denotes the complementer set of the S_i segments in S. This formula always makes use of every segment-based estimate, each of these falling into the first or the second product depending on whether it belongs to the segmentation under evaluation or not. This is why we can expect a more balanced behavior from this approximation.

Unfortunately, in practice the second factor of (7) contains too many components and cannot be efficiently evaluated. So we approximated it by considering only those elements in $\overline{S}$ that are 'near-misses' of the elements of S. The basic idea of this approach was taken from the SUMMIT system [6], but here we apply the anti-phones quite differently, as their framework is a generative one.

The phone recognition ability of our system was tested on the MTBA corpus, without the support of any language model. The percentage of phones correctly recognized (with the phone insertion rate kept at 10%) is shown in Table 3. On the OASIS-Numbers corpus isolated word recognition tests were

Sentence-Level Recognition Scores			
without anti-phones	anti-phone model Eq. (6)	anti-phone model Eq. (7)	HMM
53.44%	58.74%	61.34%	61.60%

Table 3: Phone recognition scores (% correct) on the MTBA corpus

Segmental Model	HMM
0.95%	0.80%

Table 4: Word error rates on the OASIS-Numbers corpus

	Feature Set	
Anti-Phone Model	I.	II.
No anti-phone model	67.17%	68.58%
anti-phone model Eq. (7)	72.28%	77.28%
RNN	72.39%	75.21%

Table 5: Word recognition accuracies with RNNs on the BeMe-Children corpus

conducted, and the scores can be seen in Table 4. Evidently, in both cases the segment-based model just managed to keep up with the HMM, but could not outperform it. As in phone classification it was clearly superior, we came to suspect that our estimation of the segmentation probability component was still not good enough. Two additional methods were tested to improve on this.

In the experiments described so far we artificially created anti-phone examples to enable the neural net to tell these from phonetic segments. Another option is to apply a learning algorithm that can perform 1-class learning, i.e. one that is capable of separating a class from its environment without having training examples from that environment. A neural structure that is suitable for this is the Replicator Neural Net (RNN) [7]. The concept behind it is simple enough: the input data is also used as the desired output data. Hence, by minimizing the mean square error during training, we force the net to reconstruct its training patterns. During testing the outlier patterns will have a higher reconstruction error, so it can be used as an indicator of the 'outlyingness' of a test pattern.

The RNN-based anti-phone modelling experiments were executed on a database containing recordings from children, and by using two different feature sets. The word recognition accuracies obtained are shown in Table 5. The scores show that the RNN is a viable alternative to our previous method which required the generation of a huge amount of anti-phone samples.

Another method that we experimented with was to train an ANN to estimate segment boundary probabilities on a frame-by frame basis. Using the output of this net we can significantly reduce the number of segment boundary hypotheses evaluated during recognition. This may both increase the recognition accuracy and speed up the recognition process. Figure 1 shows the output of this net and the sparse segmentation we generated from it. As can be seen in Table 6, with this method both the recognition error and the execution time went down on the Oasis-Numbers corpus (cf. Table 4).

The segmental feature set and phone classification results on the TIMIT corpus were published in [10]. The structure of the whole segment-based system along with phone classification, word and phone recognition results were published in [15] and [16]. The results with the RNN were published in [17].

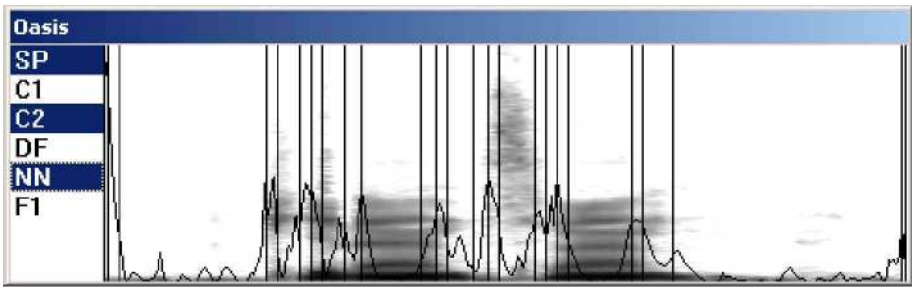

Figure 1: Sparse segmentation with the help of an ANN that learns segment boundaries

Segmentation Method	WER	Testing Time
5-frame uniform segmentation	0.95%	549,250 msec
ANN-based segmentation	0.78%	204,734 msec

Table 6: Word error rates on the OASIS-Numbers Database using ANN-based pre-segmentation

On Naive Bayes in Speech Recognition

We introduced the segmental approach with the intention of eliminating the dubious naive Bayes assumption, and we had hoped to get a superior performance from the resulting model. However, as we saw, the segmental system could only just compete with the hidden Markov model. We still believe that the HMM has a large modelling bias, but it seems that its unrealistic modelling assumptions do not significantly harm its performance. In the following we gather arguments which help us understand why this is so. Moreover, comparative experiments will be conducted with a very simple HMM and a generative segmental model. These experiments may also shed light on the comparative advantages and disadvantages of these models.

It is easy to argue against the state-conditional independence assumption of the frames in HMMs. The neighboring data frames are obviously similar because of the continuous nature of articulation, and their correlation is increased further by signal processing steps like the filtering of trajectories or the use of the delta features. Although such objections have been raised by many researchers [13], we are unaware of anyone in the speech community asking why HMMs work so well in spite of this. In the field of machine learning, however, the unexpectedly good behavior of the naive Bayes rule in classification tasks has attracted quite a lot of research. Most pertinently, it has been pointed out that in many cases naive Bayes provides optimal classification even though it incorrectly estimates the probabilities [5]. One such case is when there is full functional dependency among the features. Even when the dependency is not fully deterministic, naive Bayes classification performs nearly optimally. It is not difficult to understand why this is so: in these cases each features yields approximately the same probability estimate, so when we combine them by multiplication it is like raising one estimate to a power. The resulting estimation tends to underestimate the real probabilities, but preserves rank order and hence still leads to a correct classification.

The above arguments help explain why the HMM is able to correctly classify the segments, but we also have to explain how it solves the task of segmentation. Theoretically, it is the state transition probabilities that govern what state sequence the model goes through during operation. From this one would suspect that it is these probabilities that determine which segmentations are preferred

Phoneme Model	Classification Accuracy	Recognition Acc.	
		Unigram	Vocabulary
Frame-based, product rule	92.33%	82.05%	96.87%
Frame-based, averaging rule	78.04%	—	86.28%
Frame-based, product rule, nth root	92.33%	—	41.78%
Segmental	94.58%	46.25%	87.00%
Segmental, nth power	94.58%	57.99%	88.29%

Table 7: Classification and recognition accuracies

during decoding. Quite surprisingly, however, it has been reported by many researchers that the state transition probabilities "per se have virtually no effect on recognition performance" [4]! This means that the naive Bayes combination rule solves this problem too.

To understand how this happens, let us examine how the HMM behaves when it is allowed to evaluate all the possible segmentations. Obviously, the naive Bayes rule has a strong preference for short segments, as it multiplies small non-negative values. In practice the model is forced to fuse neighboring frames by the language model – a pronunciation dictionary or simply a large phone insertion penalty in phone recognition tasks. When it is forced to fuse neighboring frames, the Bayes rule will prefer to fuse those subsegments where one of the states provides consistently high values. If the system performs reasonably well at the frame level, then these subsegments will mostly coincide with correct phonetic segments, and those 'anti-phones' that overlap segment boundaries will be very strongly rejected.

In the experiments a 1-state HMM was compared with a simple generative segmental model. This latter used only the baseline energy features described in the previous section, and modelled these with Gaussian mixtures. This way the HMM and the segmental model differed only in their spectral component. To study the behavior of the naive Bayes rule, we tried to compensate for its bias by taking the nth root of its estimates (n being the number of frames in the segment). We also tried to introduce a similar bias into the segmental model by taking the nth power of its estimates. And out of curiosity, we also experimented with the combination of the frames by averaging them. In the tests carried out here we applied the Oasis-Numbers database.

The experimental results summarized in Table 7 confirm our suspicions. The classification results show that although the segment-based model performed the best, the naive Bayes product rule was only slightly worse. This accords with what was said about its good classification abilities. The experiments with the nth root and power indicate that the bias of naive Bayes for short segments is not detrimental at all; in fact, it even improved on the decoding abilities of the segmental model. These findings taken together show that good classification (in the segment-based model) does not necessarily mean good recognition, and that solving the segmentation problem efficiently is at least as important. While the HMM handles this issue automatically, the segment-based system has to deal with it in a special way.

Our thinking and experiments concerning the naive Bayes rule were published in [18].

The Averaging Hybrid HMM/ANN Model

Realizing the problems of the segment-based framework, we decided to return to the frame-based approach. As we still insisted on working with posteriors, our attention naturally turned towards the HMM/ANN hybrid models. Working with HMM-based systems again, we decided to focus on trying to eliminate the naive Bayes assumption when combining the frames. In analyzing the behavior of these systems we retained our segment-oriented view on the decoding process. Most importantly, we sought to identify the phone classifier component and the segment probability component of the segment-based model within the frame-based HMM/ANN hybrid. This yielded some rather interesting findings and an alternative model which we call the 'averaging HMM/ANN hybrid'.

In the HMM/ANN hybrid model [3] the component given by Eq. (3) in the HMM is replaced by

$$p(X_i|u_i) \propto \prod_{j=s_{i-1}}^{s_i-1} \frac{P(u_i|x_j)}{P(u_i)}, \tag{8}$$

where $P(u_i|x_j)$ is estimated via neural nets. With the help of Bayes' rule we can convert the left-hand side to a posterior form, and find that the *phone classifier component* $P(u_i|X_i)$ of the segment-based model (cf. Eq. (5)) in the hybrid takes the form

$$P(u_i|X_i) \approx \frac{\prod_{j=s_{i-1}}^{s_i-1} P(u_i|x_j)}{P(u_i)^{l(i)-1}}, \tag{9}$$

where $l(i)$ denotes the length of the segment. In classifier combination theory Eq. (9) is known as the *product rule* for combining class posterior estimates.

The role of the division by the class priors $P(u_i)^{l(i)-1}$ is considered controversial by the inventors of the model [3]. Hence we also did experiments where we left out this division. The resulting formulation will be referred to as the *simplified product rule*. Yet another idea was to try to take the average of the frames, which led to the *averaging rule*. These two rules are given by the equations

$$P(u_i|X_i) \approx \prod_{j=s_{i-1}}^{s_i-1} P(u_i|x_j) \quad \text{and} \quad P(u_i|X_i) \approx \frac{\sum_{j=s_{i-1}}^{s_i-1} P(u_i|x_j)}{l(i)}. \tag{10}$$

Another point worth noting is that while the averaging rule guarantees that the $P(u_i|X_i)$ estimates over the possible u_i values $\{c_1, ..., c_K\}$ add up to one, the product rules do not. Thus we introduced two further combination formulae that extend the product rule with an additional normalization step. These will be called the *normalized product rule* and the *normalized simplified product rule*, respectively.

We then performed two kinds of test to determine which rule serves as the best phone model. One of these was quite obviously to measure their phone classification performance. But from the findings with naive Bayes in the previous section we know that good classification does not necessarily mean good probability estimates. For this reason we devised another method that is hopefully more sensitive. The method is based on marginalization, more precisely on the identity $\int_x p(x)P(u|x)dx = P(u)$. We can create an estimate of the left-hand side by averaging the combination rule outputs, while the right-hand side can be estimated by simple label counting. The difference between the two can then be used as an indicator which tells us how good the combination rules are.

Combination rule	PhER	MSE
product rule	43.19%	$8.12 \cdot 10^{110}$
simplified product rule	42.44%	$7.16 \cdot 10^{-4}$
averaging rule	43.29%	$5.77 \cdot 10^{-5}$
normalized product rule	43.19%	$1.34 \cdot 10^{-4}$
normalized simplified product rule	42.44%	$5.09 \cdot 10^{-5}$

Table 8: Phone classification error rates (PhER) and the mean squared difference (MSE) between $\hat{P}(u)$ and $\frac{1}{|x|}\sum_x \hat{P}(u|x)dx$ with the various combination rules

Table 8 shows the phone classification error rates and the result of the marginalization experiment on the MTBA corpus. In phone classification the performance of each method was practically the same, while the marginalization results suggest that the normalized formulae give better probability estimates.

The next step was to examine how the combination rules behave in a phone decoding task. Only the product rule and the simplified product rule could overcome this obstacle, while the three normalized rules failed miserably. Since in classification they were not at all worse, we concluded that the reason for their failure was their inability to find the correct segmentation. Hence, it was then important to look for the *segment probability factor* of the hybrid model. As normalization destroyed the ability of the product rules to decode a phone string, we investigated the effect of normalization on the decoding process.

Knowing that the neural net guarantees that the outputs belonging to the phone classes will always add up to one, we can derive the sum of the phone class posteriors estimates of the simplified product rule. As a result, we get

$$\sum_{k=1}^{K} \hat{P}(u_i = c_k | x_{s_{i-1}}^{s_i - 1}) = 1 - \sum_{\substack{1 \leq k_{s_{i-1}}, \cdots, k_{s_i - 1} \leq K \\ \exists p, r \, : \, k_p \neq k_r}} \left(\prod_{j=s_{i-1}}^{s_i - 1} \hat{P}(u_i = c_{k_j} | x_j) \right). \qquad (11)$$

Expressed verbally, the sum on the right-hand side of (11) contains all products with mixed c_k class targets. The larger the disagreement between the frame-based experts, the larger this term becomes. Consequently, Eq. (11) can be interpreted as an estimate of $P(S_i|X_i)$, and we can say that we have found the segment probability factor of the recognizer that uses the simplified product rule: $P(S_i|X_i)$ is actually present in the combination rule, as its estimates do not add up to one, but to $P(S_i|X_i)$! That is, we can view this hybrid as if it were applying the normalized simplified product rule to estimate $P(u_i|X_i)$ and Eq. (11) to estimate $P(S_i|X_i)$.

To verify our suspicions, we tested whether Eq. (11) would enable the averaging rule to perform phone recognition and/or isolated-word recognition. The resulting compound which we called the 'averaging hybrid' performed the best in both tasks. The results are shown in Table 9 (once again using the MTBA corpus).

The idea of the averaging hybrid model was briefly outlined in [19].

Model	HMM	Prod.	Simp.Prod.	Avg.Hyb.
CORRECT	56.29%	56.82%	57.46%	57.01%
ACCURACY	46.38%	46.44%	47.71%	47.92%

Model	WER
HMM	4.80%
conventional hybrid	4.57%
averaging hybrid	3.20%

Table 9: Phone and word recognition performance of the various hybrids on the MTBA corpus

Explicit Duration Modelling in HMM/ANN Hybrids

In some languages like Finnish or Hungarian duration is a distinctive acoustic cue. But the conventional HMM framework is known to poorly model the duration information. Though the product of the state transition probabilities can be regarded as a geometric (or exponential) duration model, the exponential distribution is a very inaccurate approximation of real phone durations. Moreover, several authors have reported that the state transition values have practically no effect on the recognition scores [4]. So we decided to compare the effect of using different types of duration models within the framework of HMM/ANN hybrids. Both the conventional (product rule based) hybrid structure and the averaging hybrid were experimented with. The phone duration probability estimate was combined with $P(u_i|X_i) \cdot P(S_i|X_i)$ by multiplication after raising it to a properly tuned power α_D. In addition, the model was extended with a phone insertion penalty factor I. The optimal values for α_D and I were found by a global optimization algorithm called SNOBFIT [8].

As duration models the following possibilities were tried. In one configuration *no duration model* was used at all. This configuration served as a baseline which all the other methods could be compared to. In the next configuration a set of *exponential duration models* were applied, one model separately tuned for each phone. This corresponds to the usual way of modelling durations in HMMs. We also tried using one common, *shared exponential duration model* for all the phones, and tried extending the *exponential duration model with a minimum duration restriction* of 4 frames. Finally, we tested the *gamma duration model* that fits a gamma distribution on the duration histogram [14]. Figure 2 shows how well the various duration models approximate the duration distribution of a certain phone.

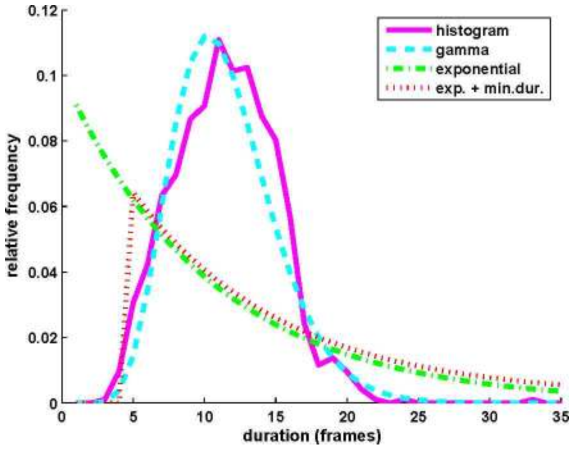

Figure 2: Fitting a duration histogram using various probability density functions

Duration Model	Conventional Hybrid			Averaging Hybrid		
	α_D	I	WER	α_D	I	WER
No duration model	–	1.511	**5.80%**	–	0.254	**4.87%**
Shared exp. dur. mod.	0.266	2.036	**5.80%**	0.934	0.806	**4.87%**
Exponential dur. mod.	0.340	3.804	**5.80%**	0.560	1.098	**4.87%**
Gamma duration model	0.382	3.311	**5.10%**	0.306	0.415	**3.94%**

Table 10: Word error rates (WER) after fine-tuning α_D and I

Without doubt, the gamma model gives by far the best estimate. In the pilot studies the minimum duration restriction was always found to improve the recognition scores. Moreover, as it also significantly speeds up the recognition process, it was turned on in all the concrete experiments.

The word recognition error rates on the MTBA corpus are shown in Table 10. The results indicate that – with properly chosen α_D and I values – the configuration with no duration model can work just as well as the configurations with the various exponential duration models. Only the gamma distribution model led to results that stand out from the rest. The gamma model brought an improvement in the performance of both the conventional and the averaging hybrid.

The experiments with the various duration models were published in [19].

Resampling-Based Training of HMM/ANN Hybrids

Throughout this dissertation artificial neural nets were applied to estimate the phone posterior probabilities – frame-based or segment-based, depending on the actual structure of the system. This is made possible by a nice theoretical proof which shows that, under proper conditions, the ANN outputs indeed coincide with the class posteriors [2]. In practice, however, the premises of the proof – an infinite amount of training data, for example – cannot be fulfilled. Hence, we will only obtain an approximation that is prone to some typical types of inaccuracies. A characteristic example of this is when the number of training examples from the various classes is significantly different. In such situations it is frequently observed that the net tends to behave inaccurately on the classes represented by only a few examples. A common solution is to present more examples to the net from the rarer classes. These methods come under the general name of 'resampling techniques' [22].

We investigated the applicability of one particular resampling method, the 'probabilistic sampling' technique [11] in the training of the ANN of HMM/ANN hybrids. This method proposes a two-step sampling scheme where in the first step a class c_k is chosen according to some probability distribution $P'(c_k)$, and in the second step a training data item is chosen from that class according to $p(x|c_k)$.

We examined how the a posteriori probability proof is affected by the choice of $P'(c_k)$. It is easy to see that, in general, a modification of the distribution of the training data invalidates the proof. However, in the case of the probabilistic sampling method there are two special cases that are of interest to us. One of them is, of course, when $P'(c_k)$ equals $P(c_k)$, the natural distribution of the classes. In this case the probabilistic sampling method coincides with the classic 'full sampling' scheme, hence the network outputs estimate $P(c_k|x)$. The other special case is when $P'(c_k) = \frac{1}{K}$, that is each class is chosen with the same probability. We call this case the 'uniform class sampling'

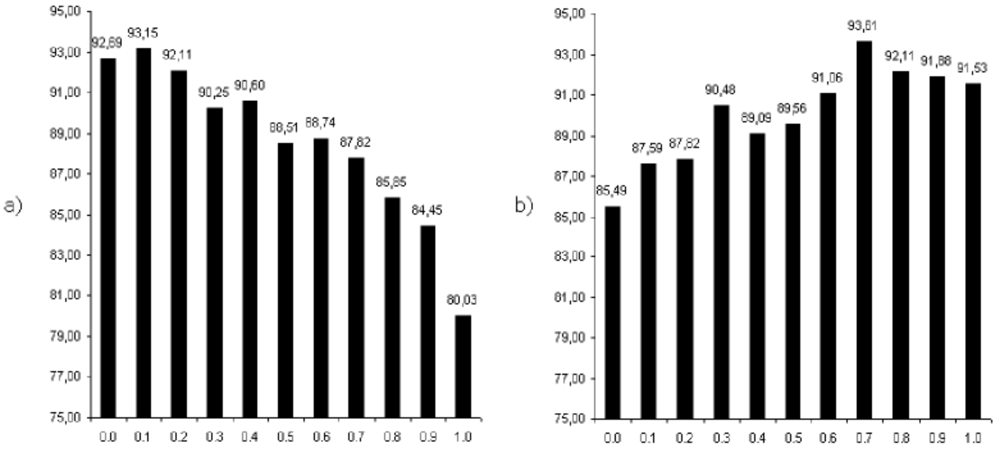

Figure 3: Word recognition accuracies (%) as a function of λ, with and without division by the priors

scheme, and we show that in this case the network outputs will be proportional to $p(x|c_k)$. Recall that the conventional HMM/ANN hybrid requires just such estimates, and in the standard case the fraction $P(c_k|x)/P(c_k)$ is used for this purpose (cf. Eq. (8)). Now we can see that the probabilistic sampling method offers an alternative solution: we can train the net using uniform class sampling, and then in the hybrid the division by $P(c_k)$ becomes unnecessary.

The proof adjusted to the 'uniform class sampling' scheme is just as sensitive to the imperfect practical conditions as the original proof was. Hence in the experiments we created a continuous transition between the two schemes by writing $P'(c_k)$ as $(1-\lambda)P(c_k)+\lambda\frac{1}{K}$, where λ was varied between 0 and 1. For each λ value the hybrid was evaluated both with a division by the priors $P(c_k)$ and without it. Figure 3 shows the word recognition accuracies obtained on the MTBA corpus, both with and without this division. The results reveal that the optima are not where the proofs predict them to be – at $\lambda = 0.0$ and $\lambda = 1.0$ – but actually somewhere in between. This example nicely demonstrates why speech recognition is more of a craft than a science.

The experiments with resampling-based training were published in [20].

Conclusions

In the author's opinion the most important result of the dissertation is not the two proposed models, but rather the insight gained with their help. The segment-based model showed that phonetic segments can be classified better via a simple and intuitive representation than by HMMs. Although the segment-based approach did not prove superior in decoding tasks, the segment-based view itself provided a new insight into what is going on in frame-based recognizers. Both the conventional HMM and the HMM/ANN hybrid were examined from a segment-based point of view, and in both cases it led us to infer that it is basically the multiplication-based combination rule that enables these models to hypothesize reasonable segmentations. After having seen that both the segmental representation and the averaging rule were the same or better at phone classification, we concluded that the principal contribution of the product rule to the decoding process lies more in its *segmentation* ability rather than its *classification* ability. The author thinks that this is definitely a surprising finding, and shows that sometimes it is worth examining an old problem from an unorthodox angle.

Summary of the Author's Contributions

In the following we summarize the results of the author by arranging them into seven thesis points.

I.) The author developed a segment-based feature set for the representation of phonetic segments. He tested this feature set on several speech corpora and in combination with various classic machine learning algorithms, and demonstrated that in most cases it results in better phone classification scores than the conventional HMM phone models.

II.) The author developed various strategies for estimating the segmentation probability component of the posterior-based segmental model, based on the concept of anti-phones. He tested the proposed modelling schemes by comparing their speech recognition performance on several speech databases.

III.) The author investigated the applicability of replicator neural networks for the estimation of the segmentation probability component of segmental models.

IV.) The author investigated how the modelling bias caused by the naive Bayes assumption influences the performance of HMM phone models. Based on the observations, he argues that this bias is such that it does not deteriorate the phone classification performance of the models and it helps them in finding the correct segmentation of the input signal. These arguments together help explain why HMMs are good at phonetic decoding while their probability estimates are quite inaccurate.

V.) The author examined the behavior of the conventional HMM/ANN hybrid model from a segment-based point of view. Based on the findings of this, he introduced a novel type of HMM/ANN hybrid which combines the frame-based posterior estimates by averaging instead of multiplication. He justified experimentally that the averaging hybrid is capable of a similar or slightly better performance than the conventional hybrid.

VI.) The author examined the efficiency of using explicit duration models in the HMM/ANN framework. He found that the gamma-distribution based duration model leads to an increased recognition performance over the conventional exponential model in both the conventional and the averaging hybrid.

VII.) The author proposed a resampling-based training scheme for the training of the neural nets used in the hybrid models. In experiments the proposed algorithm resulted in modest improvements in recognition accuracy.

The research presented in the dissertation resulted in several publications. Table 11 summarizes which publication covers which item of the thesis points.

	[10]	[15]	[16]	[17]	[18]	[19]	[20]
I.	●	●	●				
II.		●	●				
III.				●			
IV.					●		
V.						●	
VI.						●	
VII.							●

Table 11: The relation between the theses and the corresponding publications

References

[1] Austin, S., Zavaliagkos, G., Makhoul, J., Schwartz, R., Speech Recognition using Segmental Neural Nets, Proceedings of ICASSP'92, Vol. 1, pp. 625-628, 1992.

[2] Bishop C. M., Neural Networks for Pattern Recognition, Clarendon Press, 1995.

[3] Bourlard, H. A., Morgan, N.: Connectionist Speech Recognition – A Hybrid Approach, Kluwer Academic, 1994.

[4] Bourlard, H., Hermansky, H. and Morgan, N., Towards Increasing Speech Recognition Error Rates, Speech Communication, pp. 205-231, May 1996.

[5] Domingos P. and Pazzani M., On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning, Vol. 29, pp. 103-130, 1997.

[6] Glass, J. R.: A Probabilistic framework for segment-based speech recognition, Computer Speech and Language, Vol. 17, pp. 137-152, 2003.

[7] Hawkins, S., He, H. X., Williams, G. J., Baxter,R. A., Outlier Detection Using Replicator Neural Networks, Proc. DaWak'02, 2002.

[8] Huyer, W., Neumaier, A., SNOBFIT - Stable Noisy Optimization by Branch and Fit, Submitted for Publication

[9] Kleinschmidt, M., Localized Spectro-Temporal Features for Automatic Speech Recognition, Proceedings of EuroSpeech 2003, pp. 2573-2576., 2003.

[10] Kocsor, A. and Tóth, L., Kernel-Based Feature Extraction with a Speech Technology Application, IEEE Transactions on Signal Processing, Vol. 52, No. 8, pp. 2250-2263, 2004.

[11] Lawrence, S., Burns, I, Back, A, Tsoi, A. C., Giles, C. L., Neural Network Classification and Prior Class Probabilities, In: Orr et al. (eds.), Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science State-of-the-Art Surveys, Springer, pp. 299-314, 1998.

[12] Leung, H. C., Hetherington, I. L., Zue, V. W., Speech Recognition using Stochastic Segment Neural Networks, Proceedings of ICASSP'92, Vol 1., pp. 613-616, 1992.

[13] Ostendorf, M., Digalakis, V., Kimball, O. A., From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition, IEEE Trans. ASSP, Vol. 4. pp. 360-378, 1996.

[14] Pylkönnen, J., Kurimo, M., Duration Modeling Techniques for Continuous Speech Recognition, Proceedings of ICSLP' 2004, pp. 385-388, 2004.

[15] Tóth, L., Kocsor, A., Kovács, K., 2000. A Discriminative Segmental Speech Model and its Application to Hungarian Number Recognition, In: Sojka, P. et al. (eds.), Proceedings of Int. Conf. on Text, Speech and Dialogue TSD'2000, Lecture Notes in Artificial Intelligence Vol. 1902, pp. 307-313, Springer, 2000.

[16] Tóth, L., Kocsor, A., Gosztolya, G., Telephone Speech Recognition via the Combination of Knowledge Sources in a Segmental Speech Model, Acta Cybernetica, Vol. 16, pp. 643-657, 2004.

[17] Tóth, L., Gosztolya, G., Replicator Neural Networks for Outlier Modeling in Segmental Speech Recognition, In: Yin, F. et al. (eds.), Proceedings of Int. Symp. on Neural Networks ISNN'2004, Lecture Notes in Computer Science Vol. 3173, pp. 996-1001, Springer, 2004.

[18] Tóth, L., Kocsor, A., Csirik, J., On Naive Bayes in Speech Recognition, International Journal of Applied Mathematics and Computer Science, Vol. 15, No. 2, pp. 287-294, 2005.

[19] Tóth, L., Kocsor, A., Explicit Duration Modelling in HMM/ANN Hybrids, In: Matousek et al. (eds.), Proceedings of Int Conf. on Text, Speech and Dialogue TSD 2005, Lecture Notes in Artificial Intelligence Vol. 3658, pp. 310-317, Springer, 2005.

[20] Tóth, L., Kocsor, A., Training HMM/ANN Hybrid Speech Recognizers by Probabilistic Sampling, In: Duch et al. (eds.), Proceedings of Int. Conf. on Artificial Neural Networks ICANN'2005, Lecture Notes on Computer Science Vol. 3696, pp. 597-603, Springer, 2005.

[21] Verhasselt, J., Illina, I., Martens, J.-P., Gong, Y., Haton, J.-P., Assessing the importance of the segmentation probability in segment-based speech recognition, Speech Communication, Vol. 24, No. 1., pp. 51-72, 1998.

[22] Weiss, G. M., Provost, F., The Effect of Class Distribution on Classifier Learning: An Empirical Study, Tech. Report ML-TR-44, Dep. Comp. Sci., Rutgers Univ, 2002.