Explicit Duration Modelling in HMM/ANN Hybrids

No Body[1], No One[1]

[1] Institute of Nothing
South Nowhere, Neverland
{nob, noo}@nowhere.com

Abstract. In some languages like Finnish or Hungarian phone duration is a very important distinctive acoustic cue. The conventional HMM speech recognition framework, however, is known to poorly model the duration information. In this paper we compare different duration models within the framework of HMM/ANN hybrids. The tests are performed with two different hybrid models, the conventional one and the "averaging hybrid" recently proposed. Independent of the model configuration, we report that the usual exponential duration model has no detectable advantage over using no duration model at all. Similarly, applying the same fixed value for all state transition probabilities, as is usual with HMM/ANN systems, is found to have no influence on the performance. However, the practical trick of imposing a minimum duration on the phones turns out to be very useful. The key part of the paper is the introduction of the gamma distribution duration model, which proves clearly superior to the exponential one, yielding a 12-20% relative improvement in the word error rate, thus justifying the use of sophisticated duration models in speech recognition.

1 Introduction

In some languages like Finnish or Hungarian phone durations may be the only clue in discriminating certain words. Good duration modelling can therefore be an important issue. The conventional HMM speech recognition framework however does not really make use of the duration information. Though the state transition probabilities can be regarded as a geometric duration model, this model is not that effective. First, the geometric distribution is a very poor approximation of real phone durations. Second, several authors have reported that the state transition values have practically no influence on the recognition scores [2]. In this paper we examine the issue of duration modeling within the framework of HMM/ANN hybrids. Two types of hybrid models will be tested: the conventional one known from the literature, and a novel one recently proposed. In both cases we seek to answer two questions. First, we want to either prove or refute the common view that the geometric duration model is wholly ineffective. Second, we would like to know whether the replacement of the geometric model with a more sophisticated gamma distribution can improve the performance of the two hybrids.

2 A Segment-Based View of HMM/ANN Hybrids

This paper deals with the kind of HMM models where the usual Gaussian mixture component is replaced by artificial neural network (ANN) estimates. We will refer to such models as “HMM/ANN hybrids”. And, as a special case, the term “conventional HMM/ANN hybrid” here will mean the model proposed by Bourlard et al. [1]. The basic idea behind the latter is very simple: in a standard HMM, we replace the state-conditional emission likelihood estimates $\hat{P}(x_t|q_k)$ by ANN-based posterior estimates $\hat{P}(q_k|x_t)$ divided by the state priors $P(q_k)$. According to Bayes’ rule, this quotient will be proportional to the state-conditional likelihood within a scaling factor $P(x_t)$, but this factor does not influence the optimization, so the resulting system should behave like a conventional HMM.

In the following we will adopt a more general approach of the ANN-based hybrid models, where we prefer to interpret the decoding process as a search over phonetic segmentations rather than state sequences. This may be done because HMM/ANN hybrids do not use 3-state models, but have only one state per phone, so states directly correspond to phones and any state sequence naturally corresponds to a segmentation (and vice-versa). Because of this, instead of thinking in state sequences, the subsequences where the model remains in the same state can be thought of as phonetic segments; and the whole state sequence can be interpreted as a series of segments. This scheme is more general that the traditional one and will allow us to introduce a new type of hybrid model that we will call the “averaging hybrid” model. Moreover, the explicit duration models we are going to discuss can be more readily explained within this framework. However, we will also see that the conventional HMM/ANN hybrid is just a special case of this representation.

Let us now examine how the hybrid model evaluates a supposed segment. Let $X = x_1, ..., x_T$ denote the observation sequence, $U = u_1, ..., u_N$ a sequence of phonetic units over a phone set $\{q_1, ..., q_M\}$, and $S = s_0, ..., s_N$ a segmentation (given as $N+1$ segment boundary time indices).

First of all, as is usual with HMMs, we separate the acoustic and language models. Mathematically this means that we model $P(X|U)P(U)^{\alpha_L}$ instead of $P(U|X)$. The prior probability of a phone sequence, $P(U)$, is produced by the language model, and α_L is a weighting factor that is found useful in practice [4]. Here we are going to focus on the acoustic model $P(X|U)$. This factor is approximated by examining all possible state sequences or, in our jargon, segmentations S. That is,

$$P(X|U) = \sum_{S} P(X,S|U) \approx \max_{S} P(X,S|U). \qquad (1)$$

Next $P(X,S|U)$ is decomposed into segment-level scores. In our general model this decomposition looks like

$$P(X,S|U) \approx \prod_{i=1}^{N} \frac{P(u_i|x_{s_{i-1}}^{s_i-1})^{\alpha_U} \cdot P(S_i|x_{s_{i-1}}^{s_i-1})^{\alpha_S} \cdot I}{P(u_i)}, \qquad (2)$$

where $x_{s_{i-1}}^{s_i-1} = x_{s_{i-1}}, ..., x_{s_i-1}$ denotes the observation subsequence belonging to the ith segment, and $P(S_i|x_{s_{i-1}}^{s_i-1})$ can be interpreted as the probability that $x_{s_{i-1}}^{s_i-1}$ is a correct phonetic segment. The α weighting factors were simply introduced based on experience with a similar weighting factor for the language model. I is a phone insertion penalty that can be used to balance the phone insertions and deletions; again, such a factor is known to be useful in language modeling [4].

Let us examine the two main components of Eq. (2). The first, $P(u_i|x_{s_{i-1}}^{s_i-1})$, which will be referred to as P_U later on, represents the fact that each phonetic unit u_i has to be identified from $x_{s_{i-1}}^{s_i-1} = x_{s_{i-1}}, ..., x_{s_i-1}$, the signal segment mapped to it. This segment-based posterior probability can be approximated by the formula:

$$P(u_i = q_k|x_{s_{i-1}}^{s_i-1}) \approx \frac{\frac{1}{P(u_i=q_k)^{d(i)-1}} \prod_{j=s_{i-1}}^{s_i-1} \hat{P}(u_i = q_k|x_j)}{\sum_{r=1}^{M} \left[\frac{1}{P(u_i=q_r)^{d(i)-1}} \prod_{j=s_{i-1}}^{s_i-1} \hat{P}(u_i = q_r|x_j)\right]}, \qquad (3)$$

where $\hat{P}(u_i = q_r|x_j)$ are the frame-based posterior estimates and $d(i) = s_i - s_{i-1}$ is just a compact notation for the length of the segment.

In classifier combination theory Eq. (3) is known as the *product rule* and is used for obtaining an estimate of the class posteriors from the estimate of $d(i)$ independent classifiers [8]. Note that the role of the denominator is simply to normalize the estimates of the different phone classes so that they add up to one. It would not be required if the frames were truly independent. But, in fact, both theoretical arguments and experimental findings show that the frames are far from being independent. In [9] it was demonstrated that we obtain more reasonable estimates if we normalize the values and do not rely on the unrealistic independence assumption.

Alternatively, we could use the *averaging rule* of classifier combination theory:

$$P(u_i = q_k|x_{s_{i-1}}^{s_i-1}) \approx \frac{\sum_{j=s_{i-1}}^{s_i-1} \hat{P}(u_i = q_k|x_j)}{d(i)}. \qquad (4)$$

Note that in this case the estimates belonging to the various classes always add up to one, ensuring that the estimates form a correct probability distribution.

Now let us turn our attention to the other component, $P(S_i|x_{s_{i-1}}^{s_i-1})$. Its role is to compute the probability that the given segment indeed corresponds to a phone, and hence to guide the model towards finding the correct segmentation of the signal. Duration models are possible candidates because the duration information is implicitly present in $x_{s_{i-1}}^{s_i-1}$. The next section is devoted to a detailed discussion of some of the duration models that are available. Here we present an alternative, and a rather unusual interpretation of this component. This approach makes use of the frame-based posterior estimates to construct an approximation for $P(S_i|x_{s_{i-1}}^{s_i-1})$. It is based on the idea that a disagreement of the frame-based experts is likely to refer to a phonetically inhomogenious segment. Hence, it is reasonable to look for a formula that expresses the coherence of the frame-based scores. In [9] the formula

$$P(S_i|x_{s_{i-1}}^{s_i-1}) \approx \sum_{k=1}^{M} \hat{P}(u_i = q_k|x_{s_{i-1}}^{s_i-1}) \qquad (5)$$

was proposed for this purpose, based on the argument that the larger the disagreement between the frame-based experts, the smaller the value is for this expression. Consequently, it may be interpreted as a measure of incoherence of the frame-based posteriors, and can be used as an estimate for $P(S_i|x_{s_{i-1}}^{s_i-1})$. From now on we will refer to this approximation for $P(S_i|x_{s_{i-1}}^{s_i-1})$ as P_S.

Note that the incoherence of the frame-based estimates and the duration of the segment are quite different pieces of information, so it seems reasonable to make use both of them. In the experiments we will incorporate both P_S and duration models P_D in the model configurations, and they will be combined in the form $P_S^{\alpha_S} P_D^{\alpha_D}$.

3 Duration Models

No Duration Model. It has been observed by several researchers and reported in the literature that the values of the state transition probabilities have practically no effect on the recognition result [2]. Thus it is theoretically possible not to use a duration model at all. The results obtained this way can serve as a baseline for comparing the effect of the various duration models.

Exponential (Geometric) Duration Model. Hidden Markov models incorporate an implicit duration model coded by the self-transition probabilities of the states. If the self-transition probability of a state q is denoted by a_{qq}, then the probability that the models stays in state q for d steps (the duration of d frames) is $P_D(d) = (1 - a_{qq})a_{qq}^{d-1}$. This corresponds to a discrete geometric distribution, or an exponential one if we think in term of a continuous distribution. The great advantage of this exponential duration model is that it can be calculated recursively, that is $P_D(d) = P_D(d-1) \cdot a_{qq}$, so it nicely fits the dynamic programming framework of HMMs. However, in practice the duration of phones does not follow an exponential distribution. The example in Fig. 1 clearly demonstrates this fact.

The proper values for a_{qq} can be found quite easily. We only need one piece of data for this, namely the average duration for the model to stay in state q. In our one-state model the states q directly correspond to phones, so this average duration can be estimated as the mean of the phone durations over a manually segmented speech corpus. From M_q, the empirical mean of the data a_{qq} can be estimated by $a_{qq} = (M_q - 1)/M_q$ or $a_{qq} = exp(-1/M_q)$, depending on whether we are using a discrete geometric or a continuous exponential distribution.

Shared Exponential Duration Model. While in conventional HMM systems the state transition probabilities are estimated as part of the expectation maximization training procedure, in HMM/ANN systems it is common practice to use the same fixed value for all state transition probabilities [3]. It may be interpreted as if all phones had the same shared duration model. In our experiments the shared parameter value was set to 0.7.

Exponential Duration Model with Minimum Duration Restriction. If we compare the data histogram and the exponential curve fit over it in Fig. 1, we see that the largest mismatch is with small durations. A relatively simple remedy for this is to impose a minimal duration on the phones during the decoding process. For the duration model this corresponds to zeroing out the first couple of values (see Fig. 1). It is also interesting to observe that, in a 3-state model, phones are implicitly constrained to have at least 3 frames (if skipping states is forbidden). Restricting the minimal duration to 3 frames in a 1-state model will have a similar effect. Actually, in the experiments we set this value to 4 rather than 3 because this yielded slightly better results.

Gamma Distribution Duration Model. Quite evidently, the exponential duration model gives a very poor approximation of the real distribution, even with a minimum duration

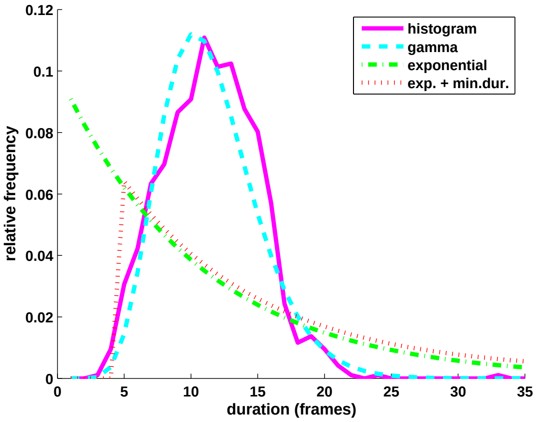

Fig. 1. Fitting a duration histogram by various pdfs.

restriction. It is natural, then, to look for another type of distribution that is only slightly more complicated, but fits the data much better. One possibility is to use the gamma distribution for this purpose. Mathematically it has the form [12]:

$$P_D(d) = \frac{(d/\beta)^{\gamma-1}e^{-d/\beta}}{\beta\Gamma(\gamma)}, \tag{6}$$

where γ is the shape parameter, β is the scale parameter, and Γ is the gamma function. The method of moments estimators of the gamma distribution are $\gamma = M_q^2/V_q$ and $\beta = V_q/M_q$, where M_q and V_q are the empirical mean and variation of the data [12].

A purely practical issue is that the gamma function cannot be computed directly but requires numerical approximations. Note, however, that it does not influence the shape of the curve but simply acts as a normalizing constant. Realizing this, we replaced it by a third parameter whose value is estimated by minimizing the mean square error between the histogram of durations and the approximation given by $P_D(d)$.

Fig. 1 shows that a gamma distribution indeed fits the data much better than an exponential distribution. The price to be paid for this is that the former cannot be computed recursively, so the usual dynamic programming decoding scheme has to be modified. This brings some additional complexity to the decoding process. Fortunately, this extra burden is manageable, because the other components (P_U and P_S) can still be computed recursively, and evaluating $P_D(d)$ for different d values is not cpu demanding. The reader should see [7] and [6] for more on how the conventional HMM or HMM/ANN structure has to be modified to incorporate explicit duration models in them.

4 Experimental Results

Database. All the results presented here were obtained using the MTBA Hungarian Telephone Speech Database [10]. This is the first Hungarian speech corpus that is publicly available and has a reasonably large size. The most important data block of the corpus contains recorded sentences that were read out loud by 500 speakers. These sentences are relatively long (40-50 phones per sentence), and were selected in such a way that together all the most frequent phone connections of Hungarian occur in them. The recordings were made via mobile and line phones, and the speakers were chosen so that their distribution corresponded to the age and gender distribution of the Hungarian population. All the sentences were manually segmented and labelled at the phone level. A set of 58 phonetic symbols was used for this purpose, but after fusing certain rarely occurring allophones, we worked with only 52 phone classes in the experiments.

For training purposes 1367 sentences were selected from the corpus. The word recognition tests described here were performed on another block of the database that contains city names. All the 500 city names (each pronounced by a different caller) were different. From the 500 recordings only 431 were employed in the tests as the rest contained significant non-stationary noise or were misread by the caller. The language model created for the words was a simple pronunciation dictionary that contained one phonetic transcript for each word and assumed that all of them had equal priors.

Preprocessing. For acoustic preprocessing we applied the Hvite module of the well-known Hidden Markov Model Toolkit (HTK) [11]. We used the most popular preprocessor configuration, that is we extracted 13 MFCC coefficients along with the corresponding delta and delta-delta values, thus obtaining the usual 39-element feature vector [11]. For recognition we used our own HMM/ANN decoder implementation, which was earlier found to have a performance similar to that of the standard HTK recognizer.

Model Configurations. The neural net used in the system contained 150 sigmoidal hidden neurons and a softmax output layer. Training was performed by conventional backpropagation. The net was trained by making use of the manual segmentation of the database, that is no embedded training was applied here (although a Viterbi-like embedded training scheme is known to be applicable to hybrid models [1]).

Two different model configurations were examined in the experiments. In our shorthand notation, the formula evaluated for each segment is

$$\frac{P_U^{\alpha_U} \cdot P_S^{\alpha_S} \cdot P_D^{\alpha_D} \cdot I}{P(u_i)}. \tag{7}$$

In the first model configuration P_U is calculated using the product rule (Eq. (3)), P_S is obtained from Eq. (5), and the duration model P_D and insertion penalty I will be varied from experiment to experiment. Both the α_U and α_S exponents will be set to 1. Notice that in this case P_S is the same as the denominator of P_U so they cancel out. Moreover, with the $P(u_i)$ in the denominator the exponent of $P(u_i)$ will become $d(i)$, the number of frames in the segment. So in practical terms what is left is the product of the frame-based probabilities, with one division by the class priors per frame. This means that this configuration is equivalent to the conventional HMM/ANN model – apart from, of course, the duration component that we are going to experiment with.

In the second configuration P_U is calculated using the averaging rule (Eq. (4)), P_S is obtained from Eq. (5), and the duration model P_D and insertion penalty I will again be varied. α_U will be set to 1, but α_S in this case will be set to 0.1, which was found to be optimal earlier [9]. We will refer to this configuration as the averaging hybrid model.

Duration Model	Model Configuration	
	Conventional	Averaging
No duration model	18.10%	34.11%
No dur. model, min.dur=4	6.04%	12.06%
Shared exponential	15.32%	10.21%
Shared exp., min.dur=4	6.96%	5.10%
Exponential	13.00%	10.21%
Exponential, min.dur=4	7.20%	9.28%

Table 1. Word error rates for various exponential model settings.

Duration Model	Conventional Hybrid			Averaging Hybrid		
	α_D	I	WER	α_D	I	WER
No duration model	–	1.5117	**5.80%**	–	0.2542	**4.87%**
Shared exp. dur. mod.	0.2667	2.0360	**5.80%**	0.9343	0.8061	**4.87%**
Exponential dur. mod.	0.3406	3.8044	**5.80%**	0.5603	1.0981	**4.87%**
Gamma duration model	0.3823	3.3117	**5.10%**	0.3069	0.4158	**3.94%**

Table 2. Word error rates (WER) after fine-tuning α_D and I.

5 Results and Discussion

In the first series of experiments we were interested in finding out how the minimum duration restriction and/or sharing a common exponential base influences the performance of the exponential duration model. In these experiments the α_D exponent and the insertion penalty I were always set to 1. Table 1 summarizes the results. From the scores it is quite apparent that the minimum duration constraint significantly improves the recognition performance (not to mention that it also dramatically decreases the run time). As regards the other question, it was surprising to see that both exponential models can be detrimental to the recognition score, and the model using the same fixed value performed better than the phone-specifically tuned one. But this was probably due to an improper choice of α_D and I (the averaging hybrid turned out to be especially sensitive to these). So the optimization of these parameters was a reasonable next step.

In the second set of experiments the weight factor α_D and insertion penalty I were fine-tuned (with the minimum duration restriction always being turned on). The optimal parameter values were found by a global optimization algorithm called SNOBFIT [5]. The resulting values along with the recognition scores are shown in Table 2. The results apparently underpin the belief that the exponential duration model brings no advantage over using no duration model at all (and, according to Table 1, with an improperly chosen exponent it can be even detrimental!). Furthermore, the practice of using one shared exponential base value instead of phone-specific ones also proved reasonable, as these models did not differ in performance. These findings seem independent of the model configuration used – conventional or averaging. In both cases only the gamma duration model was better than not applying a duration model at all. It achieved a 12-20% relative improvement in the word error rate, depending on the system configuration.

6 Conclusions

This paper investigated the feasibility of applying sophisticated duration models – in our case the gamma distribution within the framework of HMM/ANN hybrids. In addition, we were also curious to see whether the exponential duration model is indeed ineffective. Two kinds of hybrid model configurations were examined in the test, the conventional one and the recently proposed "averaging hybrid". Independent of the configuration used, we found that the exponential duration model had no detectable influence on the recognition performance. Hence the practice of replacing the phone-based self-transition probabilities by a quasi-ad hoc constant is indeed harmless – as this simplified exponential duration model is just as ineffective as the original one. On the contrary, we found that imposing a minimum duration constraint on the phonetic segments not only speeds up the decoding process, but also significantly improves the results. The other thing that yielded an improvement was the gamma duration model. Thus, altogether we are justified in saying that the exponential duration model inherent to HMM is a really poor one, and that replacing it with just a slightly more complicated model can certainly bring a modest improvement to the error rate.

Finally, let us remark that we did not discuss the differences between the conventional and the averaging hybrids because we were more interested in the duration models. But the scores clearly show the superiority of the averaging hybrid – at least, on this corpus. Moreover, during the experiments we found that the averaging model is much more tunable so, hopefully, with the introduction of new components it can be more easily improved. This is the direction we plan to take in the near future.

References

1. Bourlard, H. A., Morgan, N.: Connectionist Speech Recognition – A Hybrid Approach. Kluwer Academic (1994)
2. Bourlard, H., Hermansky, H., Morgan, N.: Towards Increasing Speech Recognition Error Rates. Speech Communication, Vol. 18., pp. 205-231, 1996.
3. Hagen, A., Morris, A.: Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR. Computer Speech and Language, Vol. 19., pp. 3-30. 2005.
4. Huang, X. D., Acero, A., Hon, H-W.: Spoken Language Processing. Prentice Hall, 2001.
5. Huyer, W., Neumaier, A.: SNOBFIT - Stable Noisy Optimization by Branch and Fit. Submitted for Publication
6. Morris, A. C., Payne, S., Bourlard, H.: Low Cost Duration Modelling for Noise Robust Speech Recognition. Proc. ICSLP' 2002, pp. 1025-1028.
7. Pylkönnen, J., Kurimo, M.: Duration Modeling Techniques for Continuous Speech Recognition. Proc. ICSLP' 2004, pp. 385-388.
8. Tax, D. M. J., van Breukelen, M., Duin, R. P. W., Kittler, J.: Combining multiple classifiers by averaging or by multiplying? Pattern Recognition Vol. 33., pp. 1475-1485, 2000.
9. Tóth, L., Kocsor, A.: Lessons from a Segment-Based Interpretation of HMM/ANN Hybrids. Submitted to Speech Communication
10. Vicsi, K, Tóth, L., Kocsor, A., Csirik, J.: MTBA – A Hungarian Telephone Speech Database. Híradástechnika, Vol. LVII, No. 8 (2002) 35- 43 (in Hungarian)
11. Young, S. et al.: The HMM Toolkit (HTK) – software and manual. http://htk.eng.cam.ac.uk
12. NIST/SEMATECH e-Handbook of Stat. Methods, http://www.itl.nist.gov/div898/handbook/