

Improvements in phone-classification accuracy from modelling duration

Philip J.B. Jackson

CVSSP, Electronic Engineering, University of Surrey, Guildford, UK. [p.jackson@surrey.ac.uk]

ABSTRACT

Durations of real speech segments do not generally exhibit exponential distributions, as modelled implicitly by the state transitions of Markov processes. Several duration models were considered for integration within a segmental-HMM recognizer: uniform, exponential, Poisson, normal, gamma and discrete. The gamma distribution fitted that measured for silence best, by an order of magnitude. Evaluations determined an appropriate weighting for duration against the acoustic models. Tests showed a reduction of 2% absolute (6+% relative) in the phone-classification error rate with gamma and discrete models; exponential ones gave approximately 1% absolute reduction, and uniform no significant improvement. These gains in performance recommend the wider application of explicit duration models.

Keywords: duration modelling, segmental HMMs.

1 INTRODUCTION

Phone durations of real speech utterances have distributions that differ markedly from those of simple Markov chains. In particular, the implicit duration model of a single state is geometric in form, whereas the mode of measured distributions can be greater than 100 ms, depending on phone identity. Even 3-state hidden Markov models (HMMs), as employed in many automatic speech recognition (ASR) systems, provide distributions that are inconsistent with duration statistics gathered from annotated speech recordings. Furthermore, annotations obtained artificially by forced alignment with an HMM-based recognizer also reveal the anomaly [1].

In studying the duration distributions of phonetic speech segments, researchers have shown that certain probability distribution functions (pdfs) provide modelling improvements. These translate into modest increases in the ASR accuracy compared to a conventional distribution, namely the exponential decay implicit in a Markov model. The durational properties of speech have been studied for many years and their importance as a cue understood since at least the

1970s [2, 3, 4, 5]. In the 1980s, there were developments in methods of modelling duration, principally for purposes of ASR: Ferguson introduced a discrete distribution to be based on counts of segment duration [6], Levinson proposed the gamma distribution and derived equations for updating the parameter estimates according to maximum likelihood [7], Russell and others considered the Poisson [8] and negative binomial [9] distributions.

Up to this point, the evidence of benefits to the recognition performance was limited, but there have been more promising results since then using the gamma distribution, in experiments on isolated words [10] and connected digits [11, 12, 13], where the normal distribution was also considered. As Burshtein has pointed out, the large number of parameters needed to represent a discrete distribution presents problems concerning their estimation from a finite amount of training data. Therefore, we would like to select a distribution function that may be described using as few parameters as necessary, which nevertheless accurately models the actual durational characteristics of speech.

This paper seeks to make a comparison of several forms of distribution within the framework of a segmental HMM (hidden segmental semi-Markov model) system [14]. The candidate forms considered here are: uniform (\mathcal{U}), exponential (\mathcal{E} , aka. geometric), Poisson (\mathcal{P}), normal (\mathcal{N} , aka. Gaussian), gamma distribution (\mathcal{G}) and discrete (\mathcal{D} , aka. Ferguson). Each has different advantages, such as computational simplicity or for comparative purposes, but can all have their parameters estimated from a set of duration statistics, which are readily calculable with our segmental recognizer, SEGVit [15].

The likelihood of a segment of duration τ may be derived from the output probability, given that the state x of the model \mathcal{M} is in state i :

$$\mathcal{L}_i(\tau) = \Pr(\mathbf{y}_1^\tau | x = i, \mathcal{M}) = d_i(\tau)^\delta \prod_{t=1}^{\tau} \mathbf{b}_i(\mathbf{y}(t)), \quad (1)$$

where \mathbf{y}_1^τ denotes the J -dimensional feature vector $\mathbf{y}(t)$ at frame t for $1 \leq t \leq \tau$, $\mathbf{b}_i(\cdot)$ the output probability, and $d_i(\tau)$ is the sampled duration distribution, which is raised to the power of the duration model scale factor δ . Here, the output probability was a diagonal-

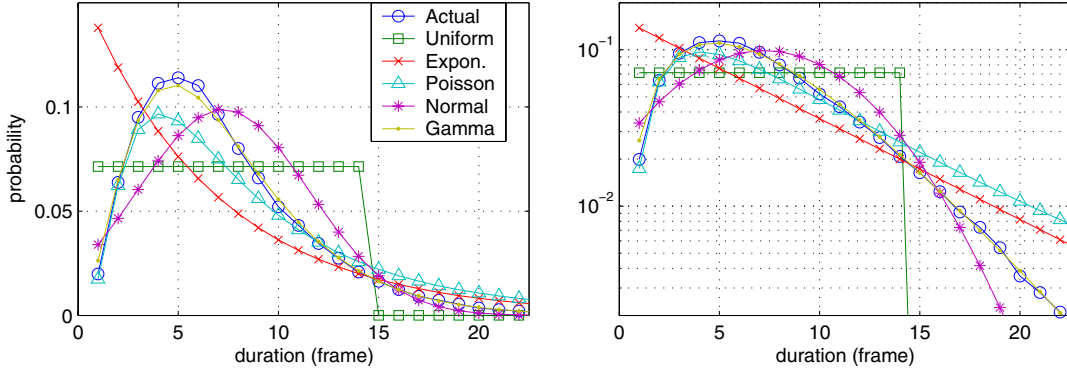


Figure 1: Forms of duration-probability distribution plotted linearly (left) and logarithmically (right), for [sil].

covariance multivariate normal distribution based on a linear-trajectory model [16], so that

$$\mathbf{b}_i(\mathbf{y}(t)) = \prod_{j=1}^J \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp - \left(\frac{(y_j(t) - f_{ij}(t))^2}{2\sigma_{ij}^2} \right), \quad (2)$$

where $f_{ij}(t) = m_{ij}(t - (\tau + 1)/2) + c_{ij}$, and m_{ij} and c_{ij} are the slope and midpoint of the trajectory for the j^{th} dimension of the i^{th} state, respectively [15].

In addition to state durations, as here, models can be learnt for phone duration, or indeed word duration [11]. Yet, the purpose of the present study was to train a set of models with differing forms of state-duration probability and to carry out comparative tests, which were done in a phoneme classification experiment on the TIMIT speech corpus.

2 METHOD

2.1 Distribution forms

The present study is based within a segmental HMM framework, which allows the inclusion of explicit duration models in a straightforward manner. Several forms of duration distribution were considered:

$$\begin{aligned} \text{Uniform} \quad d_i^{\mathcal{U}}(\tau) &= \begin{cases} 1/T_i & \text{for } 0 < \tau \leq T_i \\ 0 & \text{otherwise} \end{cases} \\ \text{Exponential} \quad d_i^{\mathcal{E}}(\tau) &= (1 - k_i) k_i^{\tau-1} \\ \text{Poisson} \quad d_i^{\mathcal{P}}(\tau) &= \frac{\mu_i^\tau}{\tau!} \exp - \mu_i \\ \text{Normal} \quad d_i^{\mathcal{N}}(\tau) &= \frac{1}{\sqrt{2\pi}\sigma_i} \exp - \left(\frac{(\tau - \mu_i)^2}{2\sigma_i^2} \right) \\ \text{Gamma} \quad d_i^{\mathcal{G}}(\tau) &= \tau^{\alpha_i-1} \exp - \left(\frac{(\beta_i \tau) \beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \right) \\ \text{Discrete} \quad d_i^{\mathcal{D}}(\tau) &= \frac{n_i(\tau)}{\sum_{t=1}^{\infty} n_i(t)} \end{aligned} \quad (3)$$

where μ_i and σ_i are the mean and standard deviation for state i respectively, the uniform pdf's maximum duration is $T_i = 2\mu_i$, the exponential time constant is $k_i = (\mu_i - 1)/\mu_i$, the gamma distribution's parameters

Error measure	\mathcal{U}	\mathcal{E}	\mathcal{P}	\mathcal{N}	\mathcal{G}
RMS ($\times 1e-3$)	8.2	10.1	8.0	5.4	0.7
Mean abs. of logs	-	0.44	1.20	0.38	0.04

Table 1: Quality of fit to [sil] for parametric duration distributions, Uniform, Exponential, Poisson, Normal and Gamma, in terms of rms. error and mean magnitude of difference in log probability.

are $\alpha_i = \mu_i^2/\sigma_i^2$ and $\beta_i = \mu_i/\sigma_i^2$, and $n_i(\tau)$ is the count of segments of duration τ in the training data.

2.2 Preliminary analysis

The TIMIT corpus of phonetically-balanced sentences was used in our experiments to provide training and test data, as 16 kHz speech files and phone annotations. Only male data were used, which includes 326 speakers from eight dialect regions of North America in the training data and 112 in the test data.

The corpus' annotations provide information about the phone durations that can be used to compile statistics from which to estimate distributions. Thus, the annotated durations of the silent sections, labelled [sil] in the database, were counted (10-ms frame rate). This illustrative example of the duration statistics is presented in figure 1, showing the distribution pattern for silence which, like those of many phones, is far from exponential. Hence, as a guide to designing the pdfs, we calculated corresponding distributions using the mean and variance of this sample, for each of the parametric forms (as in eq. 3), which are also shown in figure 1. On linear axes (left), it is clear that the models with two degrees of freedom (\mathcal{N} and \mathcal{G}) fit better than those with only one (\mathcal{U} , \mathcal{E} and \mathcal{P}), and that the gamma pdf seems best of all. These impressions are borne out by the measures of rms. error between the distributions and the duration counts in table 1, and are consistent with findings in [11, 13, 17].

However in Viterbi decoding, when comparing explanations of the observed data, the paths through

the dynamic-programming trellis multiply the segment probabilities, so it is apt to look at the multiplicative error. Thus, the mean magnitudes of the error in log probability (bottom row in tab. 1) show the same preference for \mathcal{N} and \mathcal{G} , but surprisingly the long tail of the exponential (\mathcal{E}) fitted the data almost as well as the Gaussian (\mathcal{N}), and the gamma distribution (\mathcal{G}) was an order of magnitude better than both. Considering the relatively poor performance of the Poisson and normal pdfs, they were not considered in later experiments.

2.3 Evaluation trials

Sets of phone-classification tests were conducted to evaluate the performance of the various duration distributions, using three-state acoustic models (left-to-right with skips) for 49 monophones with a single Gaussian component, as in eq. 2. A quarter of the training data were used to train the other parameters of the models (i.e., means, slopes, variances and transition probabilities), and half the remainder (i.e., another eighth) were used to perform the evaluation, with a biphone language model.

In order to train the segmental duration models, state duration statistics were gathered by SEGVit. The parameters of the duration pdfs were estimated from the training statistics, as in eq. 3, with the following exceptions:

$$\begin{aligned} \text{Uniform} \quad T_i &= \frac{1}{2} \left(15 + \frac{\mu_\phi}{3} \right) \\ \text{Exponential} \quad k_i(\tau) &= \frac{\mu'_i - 1}{\mu_i} \end{aligned} \quad (4)$$

where μ_ϕ is the mean duration for phone ϕ , and $\mu'_i = \frac{2}{3} + \frac{\mu_\phi}{3}$. As a baseline (\mathcal{B}), we set all models to have the same uniform distribution with $T_i = 15 \forall i$. Accordingly, note that \mathcal{G} and \mathcal{D} were determined by state durations, whereas \mathcal{U} and \mathcal{E} were based on those of phones. These rules were applied automatically to the estimation of the parameters for each of the 49 monophone models. However, in a few cases, they had to be manually adjusted to permit a possible state alignment to be found.¹ Duration distributions in the models were computed once for each trial, without any embedded re-estimation. To find the value of duration model scale factor that balanced the acoustic and duration models, tests were made for $\delta \in \{1, 2, 5, 8, 10, 12, 20, 50\}$, for each form.

3 RESULTS

The results in figure 2 show both the effect of varying δ and the differing performance levels of the duration functions. Since all the duration models were the same for the baseline (\mathcal{B}), the classification accuracy was not significantly affected by δ . The uniform

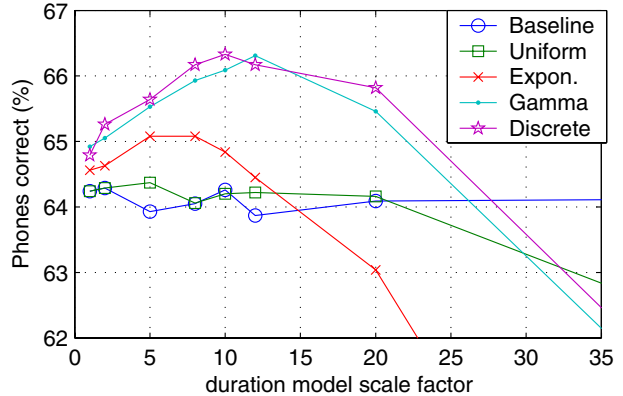


Figure 2: Phone-classification evaluations.

Distribution	\mathcal{B}	\mathcal{U}	\mathcal{E}	\mathcal{G}	\mathcal{D}
# params.	1	1	1	2	15
Eval. (best)	64.3	64.4	65.1	66.3	66.3
Eval. ($\delta = 10$)	64.3	64.2	64.8	66.1	66.3
Test ($\delta = 10$)	65.8	66.3	66.9	67.8	68.2

Table 2: Phone classification scores (% correct) for the Baseline, Uniform, Exponential, Gamma and Discrete duration distributions.

distributions provided negligible improvement but did degrade at $\delta = 50$. However, the exponential distribution, \mathcal{E} , yielded an increase of almost 1% at its peak value ($\delta = 8$) with respect to \mathcal{B} , before decreasing. In contrast, the gamma and discrete pdfs, \mathcal{G} and \mathcal{D} , continued to enhance performance up to $\delta = 10$, and demonstrated increases of 2.0% absolute above \mathcal{B} , which is comparable to those of [11, 13]. So, it seems that the models that represented the durational properties of phonemes more faithfully were the ones that performed better, and that more weight ought to be given to well-fitting duration pdfs.

Since all distributions performed close to their best at $\delta = 10$, and this value coincides with the best overall score, we used these evaluation results to fix the duration model scale factor for the final experiments, which used all the male training and test data in TIMIT respectively. The results are given in table 2, with a summary of results from the evaluation trials. While the pattern of test performance across forms is similar to the evaluations, the test results exhibit a general increase from the use of additional training data. In the case of the gamma distribution, the improvement with respect to the new baseline, \mathcal{B} , was 2.0% absolute, which corresponds to a relative reduction in the classification error rate of 6%. For the discrete distribution the increase was 2.4% absolute (7% relative).

¹These modifications are somewhat unsatisfactory, and alternative solutions are being investigated.

4 CONCLUSION

The present study was based within a segmental HMM framework, which allowed the inclusion of explicit duration models in a straightforward manner. A number of candidate distributions were considered: uniform, exponential, Poisson, normal, gamma and discrete. Analysis of the TIMIT database's phoneme annotations provided initial estimates of the duration distributions. Comparison of the candidate functions' ability to fit phone-level distributions showed the gamma distribution to outperform the other candidates by an order of magnitude.

With duration statistics derived from state occupancy during training, duration models for each of the states were estimated, which were used to investigate their effect on phone-classification accuracy. These tests determined an appropriate weighting of the duration probabilities in relation to the acoustic models, and evaluated their performance. Using these parameters, final tests were conducted. Considering the uniform distribution as a baseline, they achieved a 2.0% absolute (6% relative) reduction of the phone-classification error rate with the gamma distribution, and 2.4% (7%) with the discrete distribution, while the exponential, or geometric, distribution gave approximately 1% reduction absolute. Beyond these promising performance improvements, there are practical computational benefits to truncated duration models, which can be designed effectively according to the approach presented here. Additional experiments are needed to investigate the effects of embedded re-estimation of the duration parameters during training, and to validate our findings on a more demanding recognition task.

ACKNOWLEDGEMENTS

This research was carried out under the Balthasar project [14] (EPSRC grant GR/M87146) with Martin Russell at the University of Birmingham, UK (for further details, see <http://www.ee.surrey.ac.uk/Personal/P.Jackson/Balthasar/>). The author would like to thank him and Andrew Morris for encouragement and their helpful comments.

REFERENCES

- [1] A. Wrench and W. J. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. 5th Spch. Prod. Sem.*, Seeon, Germany, 2000, pp. 305–308.
- [2] T. P. Barnwell, "An algorithm for segment duration in a reading machine context," Tech. Rept 479, Res. Lab. of Electron., MIT, Cambridge, MA, 1971.
- [3] N. Umeda, "Vowel duration in American English," *J. Acoust. Soc. Am.*, vol. 58, pp. 434–445, 1975.
- [4] N. Umeda, "Consonant duration in American English," *J. Acoust. Soc. Am.*, vol. 61, pp. 846–858, 1977.
- [5] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 737–793, 1987.
- [6] J. D. Ferguson, "Variable duration models for speech," in *Proc. Symp. Applications HMMs Text Spch.*, Princeton, NJ, 1980, pp. 143–179.
- [7] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Comp. Speech & Lang.*, vol. 1, pp. 29–45, 1986.
- [8] M. J. Russell and R. K. Moore, "Explicit modelling of state occupancy in Hidden Markov Models for automatic speech recognition," in *Proc. IEEE-ICASSP*, 1985, vol. 1, pp. 5–8.
- [9] A. E. Cook and M. J. Russell, "Improved modelling in hidden Markov models using series-parallel configurations of states," *Proc. Inst. of Acoust.*, vol. 8, no. 7, pp. 299–321, 1986.
- [10] L. Deng, M. Lennig, and P. Mermelstein, "Use of vowel duration information in a large vocabulary word recognizer," *J. Acoust. Soc. Am.*, vol. 86, no. 2, pp. 540–548, 1989.
- [11] D. Burshtein, "Robust parametric modeling of durations in hidden Markov models," *IEEE Trans. on Spch. & Aud. Proc.*, vol. 4, no. 3, pp. 240–242, 1996.
- [12] A. C. Morris, S. Payne, and H. Bourlard, "Low cost duration modelling for noise robust speech recognition," in *Proc. Int. Conf. on Spoken Lang.*, Denver, CO, 2002, pp. 1025–1028.
- [13] N. B. Yoma and J. S. Sánchez, "MAP speaker adaptation of state duration distributions for speech recognition," *IEEE Trans. on Spch. & Aud. Proc.*, vol. 10, no. 7, pp. 443–450, 2002.
- [14] M. J. Russell, P. J. B. Jackson, N. Wilkinson, B.-H. Lo, and L. P. Wong, *Balthasar project: an integrated multiple-level statistical model for speech pattern processing*, Electron. Elec. & Comp. Eng., Univ. of Birmingham, UK, 2000.
- [15] P. J. B. Jackson and M. J. Russell, "Models of speech dynamics in a segmental-HMM recognizer using intermediate linear representations," in *Proc. Int. Conf. on Spoken Lang.*, Denver, CO, 2002, pp. 1253–1256.
- [16] M. J. Russell and W. J. Holmes, "Linear trajectory segmental HMM's," *IEEE Sig. Proc. Lett.*, vol. 4, no. 3, pp. 72–74, 1997.
- [17] T. H. Crystal and A. S. House, "Segmental durations in connected speech signals: Preliminary results," *J. Acoust. Soc. Am.*, vol. 72, no. 3, pp. 705–716, 1982.