



## Overview

Duration distributions of speech segments are not generally exponential, as HMMs imply. Using segmental models:

- various kinds of duration pdf were compared;
- evaluations determined duration-model scale factor;
- phone-classification tests were performed.<sup>3</sup>

## Theory

Probability of a segment with duration  $\tau$ , given state  $x$  is in state  $i$  of the model  $\mathcal{M}$ :

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

where  $\mathbf{y}_1^\tau$  is acoustic feature vector  $\mathbf{y}(t)$  at times  $1 \leq t \leq \tau$ ;  $\mathbf{b}_i(\cdot)$  is the output probability;  $d_i(\tau)$  is the sampled duration distribution, raised to the power of  $\delta$ , the duration model scale factor. Distributions considered were:

**Uniform**  $d_i^{\mathcal{U}}(\tau) = \begin{cases} 1/T_i & \text{for } 0 < \tau \leq T_i \\ 0 & \text{otherwise} \end{cases}$

**Exponential**  $d_i^{\mathcal{E}}(\tau) = (1 - k_i) k_i^{\tau-1}$

**Poisson**  $d_i^{\mathcal{P}}(\tau) = \frac{\mu_i^\tau}{\tau!} \exp -\mu_i$

**Normal**  $d_i^{\mathcal{N}}(\tau) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp - \left( \frac{(\tau - \mu_i)^2}{2\sigma_i^2} \right)$

**Gamma**  $d_i^{\mathcal{G}}(\tau) = \tau^{\alpha_i - 1} \exp - \left( \frac{(\beta_i \tau) \beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \right)$

**Discrete**  $d_i^{\mathcal{D}}(\tau) = \frac{n_i(\tau)}{\sum_{t=1}^{\infty} n_i(\tau)}$  (2)

where  $\mu_i$  and  $\sigma_i$  are the mean and standard deviation for  $i$ , Uniform pdf's maximum duration is  $T_i = 2\mu_i$ , Exponential pdf's time constant is  $k_i = (\mu_i - 1)/\mu_i$ , Gamma pdf's parameters 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  $\tau$  in the training data.

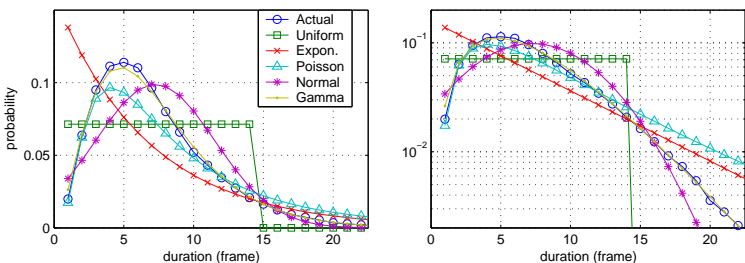


Figure 1: Forms of duration-probability distribution plotted linearly (left) and logarithmically (right), for [sil].

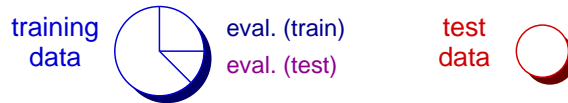
Error measure	$\mathcal{U}$	$\mathcal{E}$	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{G}$
RMS ( $\times 1e-3$ )	8.2	10.1	8.0	5.4	0.7
Mean abs. of logs	-	0.44	1.20	0.38	0.04

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.

## Experiments

Phone-classification experiments: 49 monophones (3-state segment models, left-to-right with skips, 1-mix gaussian), bigram language model,  $T_{\max} = 15$  frames.

TIMIT corpus (male only): phonetically-balanced read sentences with phonetic labels.



Duration parameters set offline by eq. 2 from training data, except: Uniform,  $T_i = (15 + \mu_\phi/3)/2$  where  $\mu_\phi$  is the mean duration for phone  $\phi$ ; Exponential,  $k_i(\tau) = (\mu'_i - 1)/\mu'_i$  where  $\mu'_i = (\mu_\phi + 2)/3$ .

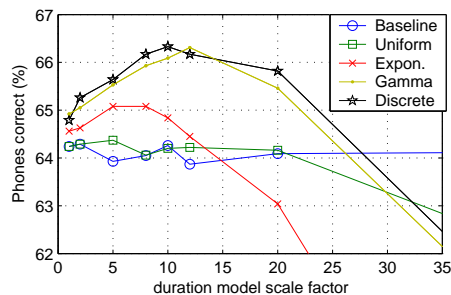


Figure 2: Phone-classification evaluations.

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.

## Conclusions

- Parametric duration models can be trivially implemented with segmental models;
- Gamma distribution most closely fitted measured silence durations;
- Exponential pdfs gave 1% absolute reduction in phone error rate, wrt. uniform pdfs;
- Gamma and discrete were best, giving 2% abs. (6% rel.), with 2 and 15 parameters per state, respectively.

## References

- [1] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Comp. Speech & Lang.*, vol. 1, pp. 29–45, 1986.
- [2] 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.
- [3] M. J. Russell, P. J. B. Jackson, N. Wilkinson, B.-H. Lo, and L. P. Wong, *Balthasar project*, Univ. of Birmingham, 2000, <http://www.ee.surrey.ac.uk/Personal/P.Jackson/Balthasar/>.

<sup>3</sup>The Balthasar project was funded by EPSRC (GR/M87146).