Gamma modelling of English segmental durations

W. N. Campbell
ATR Interpreting Telephony Research Labs

Abstract

A two-parameter Gamma function with shape and scale parameters determined by maximum likelihood estimation is shown to model the segmental durations produced by four speakers of British English better than either the normal or log-normal distributions.

1 Introduction

Segmental durations measured from readings by four speakers of British English and one speaker of American English were tested against normal, log-normal, and gamma pdfs to determine which best described their distribution.

Since the segmental durations produced by any one speaker can vary considerably between different renditions of the same utterance, the modelling of segmental duration can be considered as a stochastic process. An attempt to match any one such set of randomly observed data to a given distribution may therefore fail to be representative of the more general case. To overcome this problem, ten different sets of durations were produced for each distribution type by random generation, constrained by predetermined parameters, and an average of their correlations was compared.

2 Segmental duration pdfs

For both the synthesis and recognition of speech, it is necessary to have a satisfactory model of the distribution of segmental durations. The simplest assumption is that they fit a normal distribution.

Under the normal distribution, negative numbers can easily be generated from a low mean with a high standard deviation. This is clearly not a desirable characteristic if the numbers are to represent segmental durations. Log-scaling avoids this problem and better models the positive skew found in the distribution.

The one-parameter gamma distribution allows modelling of the shape of such a distribution, and the introduction of a second parameter, a scaling factor, enables fitting to any range of values having that distribution shape.

The bivariate gamma probability distribution function with parameters for shape and scale can be defined as

$$\Gamma(x \mid p, s) = \frac{s^{-p} x^{p-1} e^{-\frac{x}{s}}}{\Gamma(p)} \qquad (1)$$

where $\Gamma(p)$ is the gamma function, p is the shape parameter, and s is the scale parameter. Details of the maximum likelihood estimation of these parameters can be found in Campbell 1992 [2]. For high values of the shape parameter, the gamma distribution closely approximates a normal distribution, but exhibits increasing skew for smaller values.

3 Method

The Splus statistical package [1] was used on a DECstation 5000/200 to produce the sample distributions and perform the correlations. For each set of phoneme durations measured from each of five speaker's readings of 200 phonetically balanced sentences, three sets of ten arrays of equivalent length were randomly generated.

The first set was constrained to be normally distributed around the same mean and with the same standard deviation as the observed phoneme durations. The second set was similarly constrained to match the log-transformation of the phoneme durations. The third set was constrained by the shape and scale parameters of a gamma distribution, determined by maximum likelihood estimation from the observed durations for each phone type.

A correlation was performed between the original durations (sorted) and each of the ten arrays (sorted) for each set, and the mean of the correlations taken as representative of the degree of fit for each distribution.

Table 1: Fit for the distributions. Phoneme labels here are in Edinburgh University Machine Readable Phonetic Alphabet (MRPA) notation. Three phones (j, @r, and ah) produced by the US speaker are omitted from the table as they were unique to that one speaker. Better fit totals (+) for each distribution are shown after each row.

short vowels:

	@	a	e	i	o	u	uh	
normal	-	-	-	-	-	-	-	0
log	+	+	+	+	+	-	+	6
gamma	+	+	-	+	+	+	+	6

long vowels:

	@@	aa	ii	oo	ou	uu	
normal	+	-	-	-	-	-	1
log	-	-	+	+	+	+	4
gamma	+	-	+	+	+	+	5

diphthongs:

	ai	au	e@	ei	i@	oi	u@	
normal	-	-	-	-	-	-	-	0
log	+	+	-	+	-	-	-	3
gamma	+	-	-	+	-	-	-	2

stop consonants:

	p	t	k	b	d	g	jh	ch	
normal	+	+	+	-	+	-	-	-	4
log	-	-	-	-	-	-	-	-	0
gamma	+	+	+	-	+	-	-	-	4

fricative consonants:

	h	f	v	s	z	sh	zh	th	dh	
normal	-	+	-	+	-	+	-	-	-	3
log	-	-	-	+	-	-	-	-	-	1
gamma	+	+	+	+	+	+	-	-	+	7

sonorant consonants:

	m	n	ng	r	l	w	y	
normal	-	-	-	-	-	-	-	0
log	-	+	+	+	+	+	-	5
gamma	+	+	+	+	+	+	+	7

	normal	log	gamma
total + =>	8	19	31
total - =>	36	25	13

There was an average of 43 phones per speaker, with approximately 200 tokens of each, resulting in 215 averaged correlations for each distribution type. The mean values for the correlations were 0.969 for = 3.9), and the gamma showing a similar improvement on the log ($t_{428} = 3.2$). Analysis of variance for speaker-specific characteristics showed no significant differences ($F_{4,\ 210} = 1.567$), so the mean for all speakers was taken as representative for each phoneme type. A cut-off value of r = 0.98 allowed good separation of these values into two classes; 'better' fits (+) and 'worse' fits (-), as shown in Table 1.

5 Discussion

The number of 'better' results in the table above indicates that the gamma distribution does provide a closer fit to the observed data over an average of ten trials. Examination of the individual results suggests that the normal distribution is failing to model the sonorant consonants, and vowels, while the log distribution is failing to model the stop consonants.

Significant effects were found for type of phoneme by analysis of variance in both log and gamma distributions ($F_{47,\ 167} = 5.89$, and 6.16 respectively), but separate classification of vowels and consonants did not show a significant effect ($F_{1,\ 213} = 0.481$, n.s.). Separation into broad phonemic class (as in the table above) yielded significant differences at the 0.05 level with the normal ($F_{6,\ 208} = 5.922$) and log-normal distributions ($F_{6,\ 208} = 5.525$), but this effect was below the level of significance for the gamma distribution ($F_{6,\ 208} = 3.131$, n.s.), which can therefore be preferred.

No attempt has been made here to subcategorise the distributions within the phone types. In English, the great difference between durations of stressed and unstressed variants, and the extreme lengthening undergone by phrase-final segments would make this worthwhile but more sophisticated labelling of the segments would be required for such a task.

References

[1] Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988) The New S Language: A Programming Environment for Data Analysis and Graphics, AT&T Bell Laboratories, Wadsworth & Brooks/Cole Advanced books & Software, Pacific Grove California.

[2] Campbell W. N. (1992) *Prosodic Segmentation of Recorded Speech*, in PERILUS (working papers of Stockholm University, Sweden), in press.