

The Realisation of Intonational Plateaux: Effects of Foot Structure*

Rachael-Anne Knight

University of Cambridge

There is a general acknowledgement in the intonation literature that tones are often not realised as single turning points in the fundamental frequency contour, frequently occurring more as flat plateaux. House *et al's.* (1999) study of a single-speaker database demonstrates that the realisation of these plateaux co-varies with linguistic structure. The present study extends House *et al's.* analysis of the realisation of the plateau to 5 additional speakers. The earlier results are largely replicated and extended to cover an additional prosodic structure and to take account of additional measurements. Results indicate that differences found for proportional duration of the plateau are largely due to the effects of foot structure on syllable duration rather than on the absolute duration of the plateau. In addition it is suggested that plateau alignment may mark linguistic structure, specifically the difference between mono- and polysyllabic feet.

1 INTRODUCTION

Many theories of intonational phonology, especially those in the autosegmental-metrical tradition popularised by Pierrehumbert (1980), represent the intonation contour as a string of high (H) and low (L) targets. These targets are considered to be the linguistically important points of the contour whilst the transitions between them are seen as linguistically unimportant interpolations. In recent years much work has focused on describing how targets are aligned with the segmental string. In such investigations it is common to visually identify turning points associated with high and low targets and express their alignment as a duration or a percentage into some tone-bearing unit such as the syllable or foot.

Despite this general method it is often acknowledged in the literature that the search for turning points is by no means an easy one. There are many factors that make the process difficult. For example, turning points may be obscured by voiceless sections or microprosodic perturbations in the contour. Even when these are minimal, such as in sonorant stretches, the turning points may still be hard to locate as high and low parts of the contour may be sustained forming plateaux rather than sharp peaks and troughs. This phenomenon is very common (e.g. Silverman and Pierrehumbert, 1990, House and Wichmann, 1996, d'Imperio, 2002) and raises the question of whether a particular point in the plateau (such as the beginning, middle or end) or the whole plateau itself should be considered to be the speaker's real target.

When writing rules for speech synthesis this question is crucial if contours are to be expressed as a fixed number of turning points aligned with the segmental string. House *et al.* (1999) studied a medium-sized database from one male speaker of Southern British English. They attempted to reduce natural falling nuclear (H*L) contours to a small number of turning points sufficient for synthesis. Visual analysis revealed that the high tone was often realised as a plateau and informal auditory testing indicated that both ends of these plateaux needed to

*Thanks go to Francis Nolan and Sarah Hawkins at the University of Cambridge and to Jill House, Mark Hukvde and Jana Danovičová at University College London for helpful discussion of the results.

© 2002 by Rachael-Anne Knight

Luisa Astruc & Marc Richards (eds.)

Cambridge Occasional Papers in Linguistics 1:100-120.

be represented for natural sounding synthesis to result. Therefore, they decided to use alignment rules for both the beginning and end of the plateau rather than searching for a single turning point. Statistical analysis of the database revealed that the realisation of the plateau, both its duration and alignment, co-varies with linguistic structure. Specifically, when the final foot contains two syllables the plateau is longer and aligned later in the syllable than when it contains one syllable.

The fact that only a single speaker was studied is problematic for a number of reasons. Firstly, it is possible that only this individual subject produces plateaux although this seems unlikely as several studies have stated that it is often difficult to identify peaks. More likely is that other speakers produce plateaux but that these plateaux co-vary with linguistic structure in ways different from those found by House *et al.* (1999). For example, different factors may affect realisation or the same factors may have different effects. The aim of the current study is to extend the investigation of plateau realisation to five additional speakers. In order to avoid adding additional independent variables, the investigation is constrained to include only speakers of Standard Southern British English (SSBE), the same accent as that of the original speaker in House *et al.* (1999). More specifically the aim is to see if the previous results can be generalised to other speakers and also to extend the investigation by considering additional factors.

2 METHOD

2.1 Stimulus material

The experimental material consisted of 52 sentences (the majority taken from the corpus recorded by House *et al.* (1999)) designed to elicit a falling (H*L) accent on the nucleus. The final foot was controlled for the number of syllables it contained whilst final syllables were controlled for onset and coda type. The number of sentences containing each onset, coda and foot type can be seen in Table 1.

Following House *et al.* (1999) seven different onset types were distinguished. Four of these correspond to traditional phonetic categories. These are approximant, nasal, voiced obstruent and voiceless obstruent. Two types of obstruent-approximant clusters were also distinguished. One type, voiced clusters, contain a voiced obstruent and sonorant (for example /dr/) whilst the other type, voiceless clusters, contain a voiceless obstruent and devoiced sonorant (for example /tʀ/). Syllables with empty onsets (where there are no consonants before the syllable nucleus) were also included.

Four coda types were distinguished. Codas were classified either as sonorants (including approximants and nasals) or as voiced or voiceless obstruents, or as empty (where there are no consonants after the syllable nucleus).

Feet are defined according to the Abercrombian tradition (Abercrombie, 1965). Thus feet consist of one stressed syllable followed by a number of unstressed syllables. The final foot of each sentence in the experiment contained one, two or three syllables. This is an extension of the work published by House *et al.* (1999) where final feet contain only one or two syllables.

		Number of sentences
ONSET	Approximant	8
	Devoiced cluster	7
	Voiced Cluster	7
	Empty	7
	Nasal	7
	Voiceless Obstruent	7
	Voiced Obstruent	9
CODA	Empty	12
	Sonorant	13
	Voiceless Obstruent	14
	Voiced Obstruent	13
FOOT	1.00	18
	2.00	18
	3.00	16

Table 1 The number of test sentences with each onset, coda and foot type

A context was devised for each sentence in order to elicit a falling (H*L) nucleus on the target syllable without giving explicit instructions to the subject. For example:

- (1) She decided to buy the blue one. She liked it because it was glazed.

Six tokens of each of the 52 sentences were randomised (along with 6 tokens of 28 sentences used as data for a separate experiment) with the restriction that 2 tokens of any one sentence were never presented consecutively.

2.2 Subjects

Subjects were five students at the University of Cambridge. Two were male (MJJ and MAW) and three were female (RHS, CMH and SEW). Their ages ranged from 21 to 33. All were monolingual, native speakers of SSBE, none of whom reported any speech, language or hearing disorders. They were paid a small fee for their participation.

2.3 Procedure

Recording took place in a sound-treated booth using a high-quality microphone. Subjects sat in front of a 28-inch monitor on which sentence-context pairs were displayed. They were instructed to read the first sentence (the context) silently in their head and then to read the second sentence (the test sentence) out loud as if it followed the first. After each test sentence an on-screen prompt instructed subjects to begin their recording. They pressed a button allowing them to start each recording in their own time. Recordings were made directly on to the hard drive of a UNIX workstation. After each recording a waveform was displayed allowing the researcher to check the signal quality. If this was poor or the wrong intonation contour was used subjects were instructed to repeat the recording. Each sentence was recorded as many times as necessary, although usually one repetition was sufficient. Subjects were permitted to take breaks as often as they liked. On average recording sessions lasted 1.5 hours. Before the main experiment a practice recording was made of 4 sentences not included in the main experiment in order to ensure that all equipment was working satisfactorily and to familiarise subjects with the procedure.

2.4 Method of analysis

All measurements were made using ESPS Xwaves. Measurements were taken individually, by hand. Four measurements were taken from the fundamental frequency contour for each utterance. After identification of the absolute pitch peak in the nucleus, an algorithm was applied to identify the range of values that fell within 4% of the peak. 4% approximates the range of perceptual equality (Fourcin and Rosen, 1986) so all values within this range should sound to the listener to be of the same frequency. This algorithm allowed for identification of the start (SP) and end (EP) of the plateau. Durations of the accented syllable and final foot were measured. From these measurements other variables were derived. These were the absolute and relative duration of the plateau and the alignment of SP and EP in relation to the syllable and foot. The absolute duration of the plateau is calculated by measuring the duration between SP and EP. To calculate relative duration this absolute duration is expressed as a proportion of syllable or foot duration. For the alignment data, the duration between SP or EP and the beginning of the syllable is measured and this duration is expressed as a percentage of the unit (syllable or foot) in question. The measurements taken are shown in Figure 1.

Means were taken over all six repetitions of each sentence leaving 52 means for each variable for each subject. Following the methods used by House *et al.* (1999), for each subject, a univariate analysis of variance was conducted for each independent variable with fixed factors of onset type, coda type and foot type. Post-hoc comparisons for significant main effects are analysed using Tukey's HSD test. For simplification, only the results for foot type will be presented here, and alignment and relative duration are only considered in relation to the syllable.

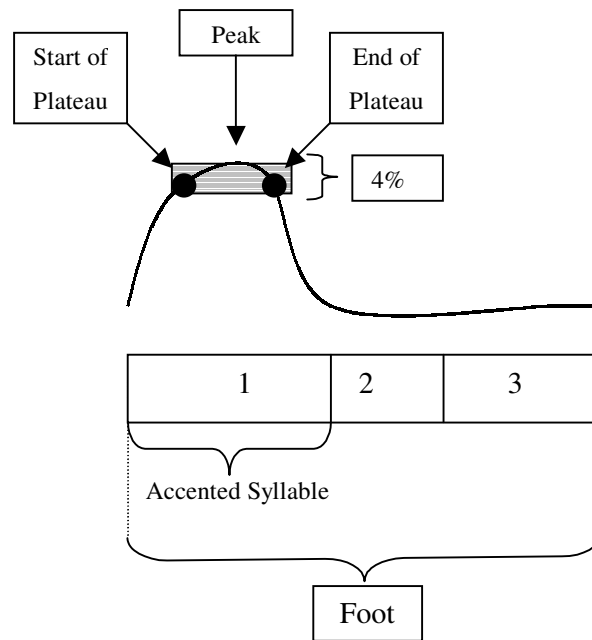


Figure 1 Measurements taken from the contour

3 RESULTS

3.1 Alignment of the start of the plateau

Foot type is not significant for CMH ($F(2,40)=2.417$, $p>0.05$). For the other subjects, as for the subject in House *et al.* (1999), there is always a trend for SP to be aligned later when there are more syllables in the foot although the exact details differ. SP is aligned earlier in the syllable in monosyllabic than in polysyllabic feet for both MJJ ($F(2,40)=20.164$, $p<0.01$) and SEW ($F(2,40)=14.447$, $p<0.01$). For these two speakers there is no difference in alignment between di- and trisyllabic feet. For MAW ($F(2,40)=6.598$, $p<0.01$) SP is aligned significantly later in trisyllabic than disyllabic feet and significantly later in disyllabic than monosyllabic feet. For RHS ($F(2,40)=19.889$, $p<0.01$) SP is earliest in monosyllabic feet and latest in trisyllabic feet whilst alignment in disyllabic feet is not significantly different from alignment in either mono- or trisyllabic feet. This is a slightly different result to the other speakers as there is a more gradient than categorical effect of foot type. These results can be seen in Figure 2.

3.2 Alignment of the end of the plateau

For every subject the number of syllables in the foot has a significant effect on the alignment of EP within the syllable (MJJ: ($F(6,40)=17.839$, $p<0.01$, MAW: ($F(6,40)=16.850$, $p<0.01$, CMH: ($F(6,40)=18.923$, $p<0.01$, SEW: ($F(6,40)=17.404$, $p<0.01$, RHS: ($F(6,40)=21.748$, $p<0.01$). In each case EP is aligned earlier in the syllable when the foot is monosyllabic than when the foot is either di- or trisyllabic. This both replicates and extends the results of House *et al.* (1999) who found a significant difference between alignment in syllables in monosyllabic and disyllabic feet. In the present results alignment is earlier in monosyllabic feet but there is no significant difference between alignment in di- and tri- syllables suggesting a categorical difference between alignment in mono- and polysyllabic feet rather than a steadily later alignment with increased syllable count. These results are shown in Figure 2.

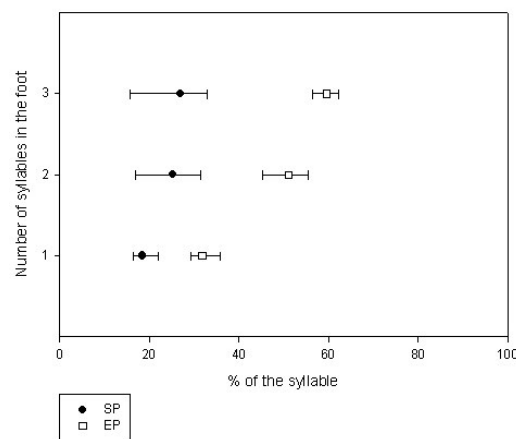


Figure 2 The alignment of the plateau in relation to the syllable as a function of the number of syllables in the foot. Circles represent the median value whilst bars show maximum and minimum values.

3.3 Absolute duration of the plateau

Foot type is a significant factor for CMH ($F(2,40) = 5.188, p < 0.00$). The plateau is shorter in monosyllabic feet than in tri- or disyllabic feet. For all other speakers foot type is not a significant factor (MJJ ($F(2,40) = 0.910, p > 0.05$), MAW ($F(2,40) = 1.395, p > 0.05$), SEW ($F(2,40) = 0.525, p > 0.05$), RHS ($F(2,40) = 0.892, p > 0.05$). These results can be seen in Figure 3.

3.4 Relative duration of the plateau

Foot type is a significant factor for every subject. For SEW ($F(2,40) = 5.192, p = 0.01$) the duration of the plateau is shorter in monosyllabic feet than in trisyllabic feet although the duration in disyllabic feet does not differ significantly from either of the other two groups. For every other subject plateaux are shorter in monosyllabic feet than in polysyllabic feet (MJJ: $F(2,40) = 4.715, p < 0.05$, MAW: $F(2,40) = 9.348, p < 0.01$, CMH, $F(2,40) = 5.192, p < 0.01$, RHS: $F(2,40) = 6.342, p < 0.01$). This mirrors the result from House *et al.* (1999) but again suggests a categorical difference between alignment in mono- and polysyllabic feet. These results can be seen by comparing the distance between the black and white circles in Figure 2

3.5 Duration of the syllable

Foot type is a significant factor for each subject. For MJJ ($F(2,40) = 9.960, p < 0.01$), SEW ($F(2,40) = 59.319, p < 0.01$) and RHS ($F(2,40) = 57.577, p < 0.01$) syllables are shortest when feet are trisyllabic, longer when they are disyllabic and longest when they are monosyllabic. Results from these subjects replicate observations by Steele (1986). However, for MAW ($F(2,40) = 34.621, p < 0.01$) and CMH ($F(2,40) = 32.841, p < 0.01$) there is no significant difference in syllable duration when feet contain three or two syllables, whilst syllables are longest in monosyllabic feet. This result can be seen in Figure 3.

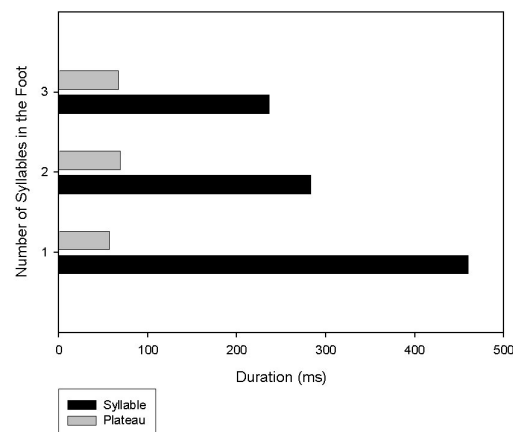


Figure 3 Duration of the syllable and plateau as a function of the number of syllables in the foot

4 DISCUSSION

The results of House *et al.* (1999) have largely been replicated and it seems that, in general, the same factors affect plateau realisation in the same way for all the speakers. The main differences occur in the results of the post-hoc tests where there is often a subtly different grouping of levels although the general trend is the same.

The results for alignment replicate those of House *et al.* (1999) in that both ends of the plateau are later when there are more syllables in the foot. In addition the current results extend those of the earlier study by showing that in general there is a categorical difference between alignment in mono- and polysyllabic feet rather than a gradient difference between feet containing increasingly more syllables. These details of plateau alignment can be related to the differences in peak alignment found when the syllable is lengthened by prosodic factors. Silverman and Pierrehumbert (1990) and Steele (1986) both demonstrate that when syllables are lengthened by prosodic means nuclear and prenuclear peaks are aligned earlier within the lengthened unit. This is the case for plateau alignment in the present study. When a syllable is the only one in a foot it is lengthened by utterance final lengthening and is significantly longer than syllables that are followed by unstressed syllables. In the same environment plateaux are significantly earlier.

Results for the proportional duration of the plateau also replicate those of House *et al.* (1999) in that plateaux take up more of the syllable in longer feet than in shorter feet. The present results also extend these findings. Firstly, mirroring the findings for alignment there appears to be a categorical difference between mono- and polysyllabic feet. Secondly, measurements for the absolute duration of the plateau indicate that it is not affected by the structure of the foot. This suggests that the differences found for proportional duration are largely due to prosodic effects on the duration of the syllable rather than on the absolute duration of the plateau itself.

It is possible that prosodically induced alignment differences may be used to signal the upcoming prosodic structure of the utterance. So, for example, an earlier plateau may indicate to the listener that the foot is monosyllabic. This information could be used by the processes responsible for spoken word recognition. It has often been suggested that prosody may play a role in spoken word recognition. Strong syllables, for example, may help the listener to segment the speech stream (e.g. Cutler and Butterfield, 1992) or may assist with lexical access (Friedrich *et al.*, 2002). The findings from the current experiment suggest that subtle differences in F0 alignment could potentially be used in a similar way.

It is not entirely clear from these results which point of the plateau is likely to be the speaker's real target or be most useful in word recognition, however it should be noted that the alignment of EP co-varies with structure more consistently across speakers than the alignment of SP. In addition Knight (2002) demonstrates that the end of the plateau is unaffected by the non-structural variable pitch span whereas the peak and start of plateau occur later in the syllable in wider pitch spans. It therefore seems likely that EP is the speaker's real target in production and is potentially the most reliable indicator of linguistic structure in perception.

5 REFERENCES

- Abercrombie, D., 1965. *Studies in Phonetics and Linguistics*. London: OUP
- Bel, B. and I. Marlien (eds.) (2002) *Proceedings of the Speech Prosody 2002 conference, Aix-en-Provence*
- Cutler, A. and S. Butterfield (1992) "Rhythmic cues to speech segmentation: Evidence from juncture misperception", *Journal of Memory and Language*, 31, 218-236
- D'Imperio, M. (2002) Language Specific and Universal Constraints on Tonal Alignment: The nature of Targets and "Anchors", in B. Bel and I. Marlien (eds.), 101-106
- Friedrich, C., S. Kotz, A. Friederici, and K. Alter,. (2002) "Pitch contour guides spoken word recognition" in Bernard Bel and Isabelle Marlien (eds.) 311-314
- House, J., J. Dancovičová and M. Huckvde. (1999) "Intonation Modelling in ProSynth". *Proc. XIV ICPhS Vol 3, University of California, Berkeley, CA*, 2343-2346
- House, J. and A. Wichmann. 1996 "Investigating peak timing in naturally-occurring speech: from segmental constraints to discourse structure". *Speech, Hearing and Language: work in progress, Volume 9*, 99-117. University College London
- Knight, R-A. (2002). "The influence of pitch span on intonational plateaux" in B. Bel and I. Marlien (eds.), 439-442
- Pierrehumbert, J. (1980) *The Phonology and Phonetics of English Intonation*. Michigan: MIT Press
- Rosen, S. and A. Fourcin. 1986. "Frequency selectivity and the perception of speech". In Moore, B.(ed.) 1986, *Frequency Selectivity in Hearing*. London: Academic Press
- Silverman, K. and J. Pierrehumbert. (1990). "The timing of prenuclear high accents in English". In Kingston, J. and M. Beckman. (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, Cambridge: CUP 72-106
- Steele, S. (1986) "Nuclear accent F0 peak location: effects of rate, vowel, and number of following syllables". *Journal of the Acoustical Society, Supplement 1*, 80; s51

Rachael-Anne Knight

*Linguistics Department
Sidgwick Avenue
University of Cambridge
Cambridge
CB3 9DA
United Kingdom*

*rachaelanne@cantab.net
<http://www.rachaelanne.co.uk>*