Should rhythm metrics take account of fundamental frequency?*

Ruth Cumming (née Galloway)

Department of Linguistics, University of Cambridge

Research on rhythm has tended to concentrate on duration as a measurable acoustic parameter, although other factors such as f0 and intensity may contribute to rhythm perception. Some studies (not specifically concerned with rhythm) have provided evidence for a perceived lengthening effect of dynamic f0 e.g. Lehiste (1976), although others failed to replicate this finding with listeners from other language backgrounds e.g. van Dommelen (1993). The present study extends previous research by investigating listeners from two rhythmically distinct language backgrounds, and by using both linguistic and non-linguistic stimuli. In a two-alternative forced choice decision task, subjects in both language groups heard stimuli with dynamic f0 as longer than those of equal duration with a level f0. Therefore, such an effect of dynamic f0 may not be dependent on the rhythmic properties of the listeners' native language. The results are discussed with reference to the implications this has for duration-based rhythm research.

1 INTRODUCTION

1.1 Rhythm and duration

The latest decade of empirical research on speech rhythm has primarily concerned duration. This is evident within the contributions to the forthcoming special issue of *Phonetica* (2009) on empirical approaches to speech rhythm. For example, the now popular rhythm metrics such as %V, ΔC, ΔV (Ramus *et al.* 1999), various Pairwise Variability Indices (PVIs) (e.g. Grabe and Low 2002), and others, e.g. VarcoV (Dellwo 2006), CCI (Bertinetto and Bertini 2008) etc., have all measured duration of units in speech. This fascination with duration is perhaps not surprising, since the notion of 'timing' has long been deemed essential in the definition of rhythm: from Ancient Greek philosopher Aristoxenus (Adams 1979: 10), through to early 20th century psychologists like Bolton and Edwards (Wallin 1911: 108), and to a canonical reference in speech rhythm research, Pike (1945).

Nevertheless, some researchers have considered factors other than duration. The psychologist Wallin (1911) believed pitch was essential in rhythm perception, and his contemporary Woodrow (1911) thought duration, pitch and intensity all had important, interacting roles to play (see Adams (1979: 9-57) for a review of the 'temporal' versus 'accentual' views of rhythm). In the context of linguistics, Dauer (1983, 1987) asserted the importance of vowel reduction and lexical stress in rhythm; these involve spectral properties, f0, intensity and duration. Low (1998), the first to use the PVI, took not only duration, but amplitude and spectral dispersion measures too, since these may all contribute to rhythm perception.

* I would like to thank Francis Nolan, Stephan Schmid, and Antje Heinrich for their help in the conception, running and analysis of the experiment (respectively).

1.2 Does a dynamic f0 increase perceived duration?

This question seems to have first been asked in the 1970s. English speakers were found to perceive a synthetic vowel with a dynamic f0 as longer than one with a level f0 when both were of equal physical duration (Lehiste 1976, Pisoni 1976). This effect also occurred with non-speech stimuli (Wang *et al.* 1976). More recently, Yu (2006) asked English speakers to rate synthetic /pa/ monosyllables on a 7-point duration scale: dynamic stimuli were generally perceived as longer than level stimuli.

However, similar experiments with native speakers of other languages have failed to replicate this finding. Rosen (1977a) presented Swedish listeners with synthetic vowels, but only when the second vowel had a dynamic f0 did this lengthening effect occur. Rosen (1977b) played /ɛt/ stimuli with various durations and f0 contours to more Swedish listeners, who indicated whether they heard *ett* ([ɛt], one) or *ät* ([ɛːt], eat); there was no evidence for a perceived lengthening effect of dynamic f0. Likewise van Dommelen (1991, 1993) asked German listeners to indicate which word they heard out of various minimal pairs involving the /a/-/aː/ distinction e.g. *walle-Wale* /valə/-/vaːlə/, when stimuli had a falling or level f0. In some monosyllabic stimuli falling f0 led to more 'long' judgments, but in disyllabic stimuli falling f0 consistently increased the number of 'short' judgments. It was concluded 'that Lehiste (1976), through the use of isolated vowels as speech material, happened to find the exception rather than the rule' (van Dommelen 1993: 383-4). However, neither Rosen (1977b) nor van Dommelen (1991, 1993) discuss vowel quality, which may have affected their results. Swedish [ɛː] is more central and [ɛ] is more front (IPA Handbook 1999). German [a] may be higher, more lax, or indeed more front than [aː] (see Wiese (2000: 21-2) for discussion).

Lehnert-LeHouillier (2007) suggested that English speakers might not react in the same way as Swedish and German speakers because English does not have phonemic vowel length contrasts[1]. She tested speakers of Thai, German and Japanese (all of which have vowel length contrasts) and Spanish (which does not), using nonsense /tV/ monosyllables, with a falling or level f0. A perceived lengthening effect of dynamic f0 only occurred in Japanese listeners. Stimuli were manipulated from natural tokens produced by an Estonian speaker, a language with unaspirated /t/, which may have affected results, since the typical VOT of /t/ in the four test languages is diverse: aspirated in German, unaspirated in Spanish, both aspirated and unaspirated in Thai, and moderately aspirated in Japanese (IPA Handbook 1999).

1.3 The present experiment

In summary, then, dynamic f0 seems to increase perceived duration only in certain cases. Given these previous mixed findings, the present study aimed to (i) extend previous research, and (ii) discuss the implications for duration-based rhythm research (cf. Lehiste 1976, Rosen 1977a), if indeed f0 was found to interact with duration in perception.

Firstly, unlike in previous studies, two groups of listeners with a different native language were subjected to both linguistic and non-linguistic stimuli. The

[1] Since quality differences exist in Swedish and German, arguably they are similar to e.g. English *bead* versus *bid* /biːd/-/bɪd/.

languages investigated here were French (henceforth Fr) and Swiss German (henceforth SG). The linguistic stimuli were /si/ monosyllables, because this is a frequent word – Fr *if* or *yes* (in response to a negative question), and SG *she* or *they*; arguably this was a more speech-related task than listening to isolated vowels. The phonetic properties of the voiceless fricative were particularly desirable: a plosive would entail the issue that different oral and laryngeal timing patterns occur in the realization of Fr and SG plosives (Galloway 2007), which may affect subjects' perception of duration; a voiced segment would require a decision on whether the f0 movement should occur over the whole syllable, or just the vowel. The non-linguistic stimuli were 'buzzes' with no meaning. Both non-linguistic and linguistic stimuli were manipulated with identical f0 contours, thus a direct comparison of the two conditions was possible. Secondly, several types of f0 change were included here, to investigate the potential effects of direction, timing and excursion of f0 change, which may affect perception of pitch and duration (Lehiste 1976, House 1990).

1.4 French and Swiss German prosody

There are two reasons behind the choice of Fr and SG. Firstly, to my knowledge, neither has previously been subject to investigation in this context. Most previous studies tested speakers of 'standard' Germanic languages. 'Swiss German' is not one standard language, rather a group of Alemannic dialects spoken within the political borders of Switzerland; Fr is a Romance language.

Secondly, since Fr and SG are rhythmically different they are well suited to testing the potential effect of native language on perception of pitch and duration. Fr has consistently given relatively low PVI and ΔV/ΔC values (Galloway 2007, Grabe and Low 2002, Ramus *et al.* 1999, White and Mattys 2007), consistent with the theory that it lies at the 'syllable-timed' end of a rhythm-type continuum (Grabe and Low 2002). SG, although less investigated, has given relatively high PVI and ΔV/ΔC values (Galloway 2007, Schmid 2001), consistent with SG being at the 'stress-timed' end of the putative rhythmic continuum.

1.5 Hypothesis

Lehiste (1976) and Lehnert-LeHouillier (2007) both suggest, from their results, that a lengthening effect of dynamic f0 might occur only in listeners whose native language associates pitch changes with increased duration e.g. English stress. If this is true (although it might not be), the following predictions could be made.

SG: It is predicted that SG listeners will generally perceive stimuli with dynamic f0 as longer than equally long stimuli with level f0, because in SG large f0 changes and increased duration are both associated with stressed syllables (Häsler *et al.* 2005, Siebenhaar *et al.* 2004). However, experimental data concerning SG stress correlates is relatively limited and only from production (not perception) studies. Furthermore, studies similar to the present one, on listeners of other Germanic languages (apart from English) have not found a perceived lengthening effect of dynamic f0.

Fr: More research has been conducted on Fr than on SG stress, but the results appear quite confusing. In production studies, some claim that duration is the cue to Fr prominence e.g. Delattre (1966), whereas others claim it is pitch e.g. Parmenter and Blanc (1933). Benguerel (1971, 1973) measured several correlates of prominent Fr syllables, and concluded that duration and f0 were significant in

'phrase-final' and 'emphatic' stress respectively. In perception studies, Rigault (1962) asked native Fr speakers to label the most prominent syllable in disyllabic words, and found f0 to be the most significant cue, whereas a similar experiment by Mertens (1991), but which instead used short extracts of Fr, found duration to be most significant. Different types of Fr stress possibly have different cues: duration for final/phrasal stress versus f0 for emphatic stress (cf. Di Cristo and Hirst 1997, Vaissière 1991), therefore Fr-speaking listeners might not associate increased duration with f0 changes. It is predicted that they will not generally perceive stimuli with dynamic f0 as longer than equally long stimuli with level f0, and their responses will be around the level of chance.

Cross-linguistic comparison: If this perceived lengthening effect of dynamic f0 depends on native language, SG and Fr responses are predicted to differ significantly.

2 METHOD

2.1 Subjects

Two groups of listeners were recruited, all of whom reported normal hearing, and were offered a small reimbursement for their time. Fourteen native SG speakers were tested in Zürich, most of whom speak the Zürich dialect. All were students at the University of Zürich or *die Eidgenössische Technische Hochschule Zürich* (The Federal Technical College of Zürich). They ranged from 20 to 37 years old (mean = 25 years); there were more female than male volunteers.

Eight Fr speakers were tested in Zürich, and eight in Cambridge. Those tested in Switzerland were native speakers of Swiss Fr, and those tested in Cambridge were native speakers from France. According to Grosjean *et al*. (2007: 2) the two varieties are generally very similar, although differ in some lexical and phonological respects: Swiss Fr supposedly has word-final vowel length contrasts e.g. /i/-/iː/, /u/-/uː/ (Grosjean *et al.* 2007, Andreassen 2006). Grosjean *et al*. (2007: 2-3) claim (without reference) that Swiss Fr 'shows more pitch movement on penultimate syllables in phonological phrases than Parisian French'.

Since the differences between these two varieties of Fr could affect listeners' judgments, overall responses from the two Fr groups were compared statistically before the main analysis of results. However, since they were not statistically different, these sixteen subjects were treated as one group in the subsequent main analysis. The age range was 18 to 36 years (mean = 26 years). Those recruited in Zürich were mainly young working professionals, and those in Cambridge were university students or staff; there were about half male and half female volunteers.

2.2 Stimuli

The buzzes were created using the synthesize function in *Praat*, which creates a pulse train which is then run though a series of filters representing five formants. Duration and f0 were manipulated on each buzz to produce 36 different stimuli (3 durations x 12 f0 contours – table 1).

	Level (L)	Falling (F) 200-100 Hz	Rising (R) 100-200 Hz	Complex (C)
1	100 Hz			100-200-100 Hz
2	200 Hz			100-150-100 Hz
3				200-100-200 Hz
4				200-150-200 Hz
Durations for each f0 contour shown above		250ms 375ms 500ms		

Table 1 – f0 contours and durations of stimuli

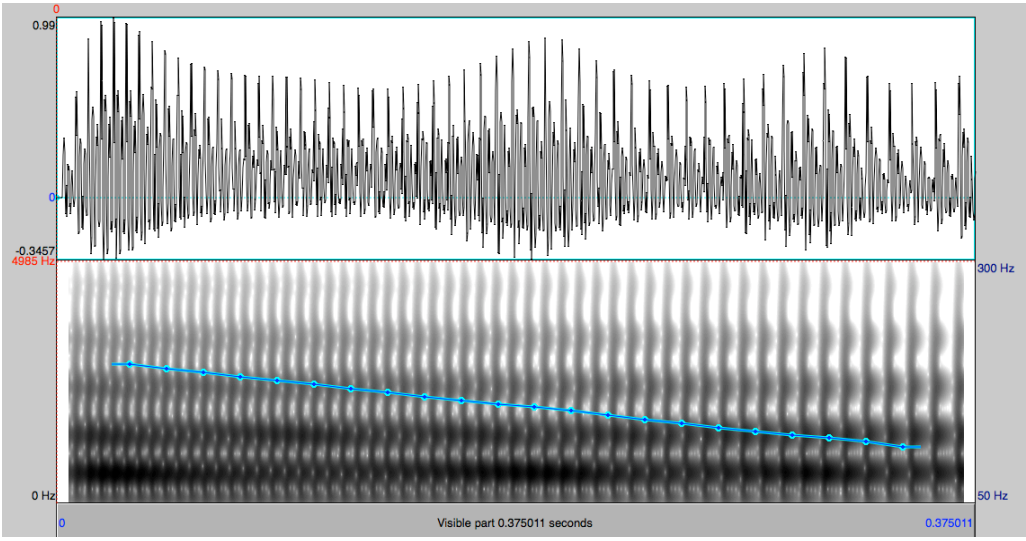

Figure 1 – Example: Spectrogram and waveform of a buzz stimulus (F1, 375 ms)

The /si/ monosyllables were manipulated in *Praat* from a single natural token produced by a phonetically-trained, male, native British English speaker articulating a series of monosyllables beginning with /s/ and ending in a cardinal vowel. The typical realisation of /i/ in each language is not so far from a cardinal production that these /si/'s would be unrecognisable to the subjects. The recording took place in a soundproof studio using a *Marantz* PMD670 solid-state recorder and a low noise condenser *Sennheiser* MKH40P48 microphone with a cardiod frequency response. The recording mode was set to 16 bit linear PCM, with a 44.1 kHz sample rate, and the data file was saved as .wav format. The file was subsequently transferred onto a *MacBook* (Mac OS X v10.4) via a USB cable, and displayed in *Praat*.

One /si/ monosyllable with a duration of 500 ms was selected for manipulation into 36 different stimuli with durations and f0 contours identical to the buzzes – see table 1 (f0 change naturally only occurred during the vowel). Firstly f0 was manipulated using *Praat*'s *PitchTier* function, which allows an exact f0 contour to be specified and added to a sound file, which is then resynthesized with that f0. Then the duration of each new (f0-manipulated) /si/ was changed; two *Praat* scripts were

written and run, to shorten each /si/ during resynthesis, one to 75% and one to 50%, consonant and vowel by the same proportion. The original was 206+294=500ms (C+V), therefore one resynthesis was 155+220=375 ms, and one was 103+147=250 ms.

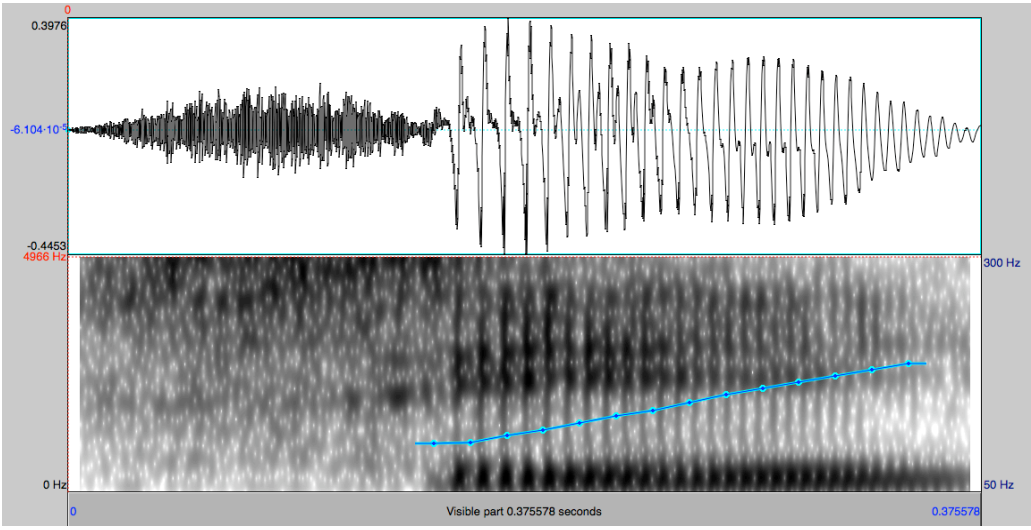

Figure 2 – Example: Spectrogram and waveform of a /si/ stimulus (R1, 375ms)

In each trial, subjects heard a pair of stimuli (inter-stimulus interval = 800 ms). Details of trials, which were completely randomized, are given in table 2.

Type		1st stimulus	2nd stimulus	Number of trials
1	**Test**: Are stimuli with dynamic f0 judged as longer than stimuli with level f0? (counterbalanced order)	L1	F(1-3) R(1-3) C(1-4)	**30** (10 dynamic f0 contours x 3 durations)
		F(1-3) R(1-3) C(1-4)	L1	**30** (10 dynamic f0 contours x 3 durations)
2	Are stimuli with high f0 judged as longer than stimuli with low f0, or vice versa? (counterbalanced order)	L1	L2	**3** (1 x 3 durations)
		L2	L1	**3** (1 x 3 durations)
3	**Identical**	L1	L1	**3** (1 x 3 durations)
		L2	L2	**3** (1 x 3 durations)
4	**Fillers**	L1 – 250 ms	L1 – 375 ms	**36** (12 pairs x 3 occurrences)
		L1 – 250 ms	L1 – 500 ms	
		L1 – 375 ms	L1 – 250 ms	
		L1 – 375 ms	L1 – 500 ms	
		L1 – 500 ms	L1 – 250 ms	
		L1 – 500 ms	L1 – 375 ms	
		L2 – 250 ms	L2 – 375 ms	
		L2 – 250 ms	L2 – 500 ms	
		L2 – 375 ms	L2 – 250 ms	
		L2 – 375 ms	L2 – 500 ms	
		L2 – 500 ms	L2 – 250 ms	
		L2 – 500 ms	L2 – 375 ms	
	Total (per section)			**108**
	Grand total			**216**

Table 2 – Summary of stimuli pairs (codes refer to those in table 1)

Type 1 trials (shaded) are the most interesting, since they test the hypothesis. The order of stimuli was counterbalanced, since Rosen (1977a) previously found an ordering effect. Responses to the type 2 trials (one high level, and one low level stimulus) are discussed elsewhere (Cumming in preparation). Type 3 trials were controls with no f0 movement, which tested whether subjects were more likely to judge the first or the second stimulus as longer. Type 4 trials were 'fillers', which comprised two stimuli with perceptibly different durations but the same level f0. These reassured subjects that there were some 'easier' trials, to prevent them becoming bored. They were expected to judge the longer stimulus as longer, thus it tested how accurately they completed the task. In the pilot, each filler was heard only twice, but this was amended to three times, after a few subjects reported suspicion that several stimuli were equal in duration.

2.3 Equipment, procedure and analysis

All subjects sat in a sound-attenuated booth, either in Zürich University's or Cambridge University's Phonetics Laboratory, and listened through binaural headphones (Zürich, *Sony* MDR-V600; Cambridge, *Sennheiser* HD520). Before testing began, they were given printed instructions, and could ask for clarification. It

was explained that they would hear many pairs of sounds, and that the question asked of them (which also appeared on screen for each trial) was: 'Which sound was longer – sound 1 or sound 2?' Their task was to click one of the two on-screen buttons labelled '1 was longer' or '2 was longer'. They were warned that some trials were harder than others, but it was a forced choice task with no abstentions permitted. All instructions and on-screen text were written in the subjects' native language: Fr for Fr speakers, and standard German for SGs.

The experiment was run in *Praat*, and comprised two sections (buzzes and /si/'s) with a break in between. The order of sections was counterbalanced across subjects, which seemed important since the results of the pilot experiment (also counterbalanced) showed a statistically significant effect of order: dynamic stimuli were perceived as longer than level stimuli significantly more often by subjects who heard the /si/ stimuli first. For each section, subjects completed 10 practice trials, which included a range of stimuli (various durations, pair orders and f0 contours) from the real experiment before it began. They were also allowed to ask for clarification after the first practice session, although none did. There were 236 trials in total: (10 x practice) + (108 x main) x 2 sections. An on-screen prompt encouraged subjects to take a short break after every 20 trials. Each trial began after subjects had clicked to begin, or to respond to the previous trial. After this click, 800 ms of silence followed, then the first stimulus; in the pilot, there was only 500 ms of silence, but some subjects indicated that they needed longer to focus on the next trial. Only one listening was allowed, but response time was unlimited. The whole experiment lasted approximately 25 minutes.

Responses were recorded in *Praat* and transferred to *Excel*. For each subject, counts of his/her responses to the various trials were made. Firstly 'filler' responses were counted; all subjects had responded accurately enough so none were rejected. The percentage of correct filler responses was as follows: SG buzzes, $\bar{x}$ = 96.6, s = 2.71; SG /si/, $\bar{x}$ = 97.6, s = 4.05; Fr buzzes, $\bar{x}$ = 93.7, s = 4.42; Fr /si/, $\bar{x}$ = 95.8, s = 4.12.

Then all data were explored graphically, and tests of normality (Shapiro-Wilk) and homogeneity of variance (Levene) were conducted. Further statistical analysis consisted of two stages addressing the following:

1. Did subjects in each group respond 'dynamic stimulus is longer than level stimulus' (henceforth 'D > L') at a level significantly above chance?
2. Do any other variables have an effect?

ANOVAs were computed with *SPSS* and binomial tests with *Excel*.

3 RESULTS

Firstly, do subjects generally perceive dynamic stimuli as longer than level stimuli, and is there a difference between language groups? (see table 3, figure 3)

	'D > L' responses (number, mean percentage and standard deviation from *k* subjects)	
Language Group	*Buzzes*	*/si/ monosyllables*
Swiss German ($k = 14$)	$n = 513$ ($N = 840$) $\bar{x} = 61.1$, $s = 13.3$	$n = 543$ ($N = 840$) $\bar{x} = 64.6$, $s = 14.3$
French ($k = 16$)	$n = 668$ ($N = 960$) $\bar{x} = 69.6$, $s = 10.3$	$n = 693$ ($N = 960$) $\bar{x} = 72.2$, $s = 12.5$

'D > L' = 'dynamic stimulus is perceived as longer than level stimulus'; k = number of subjects per group; n = number of 'D > L' responses; N = total number of trials

Table 3 – Summary of 'D > L' responses

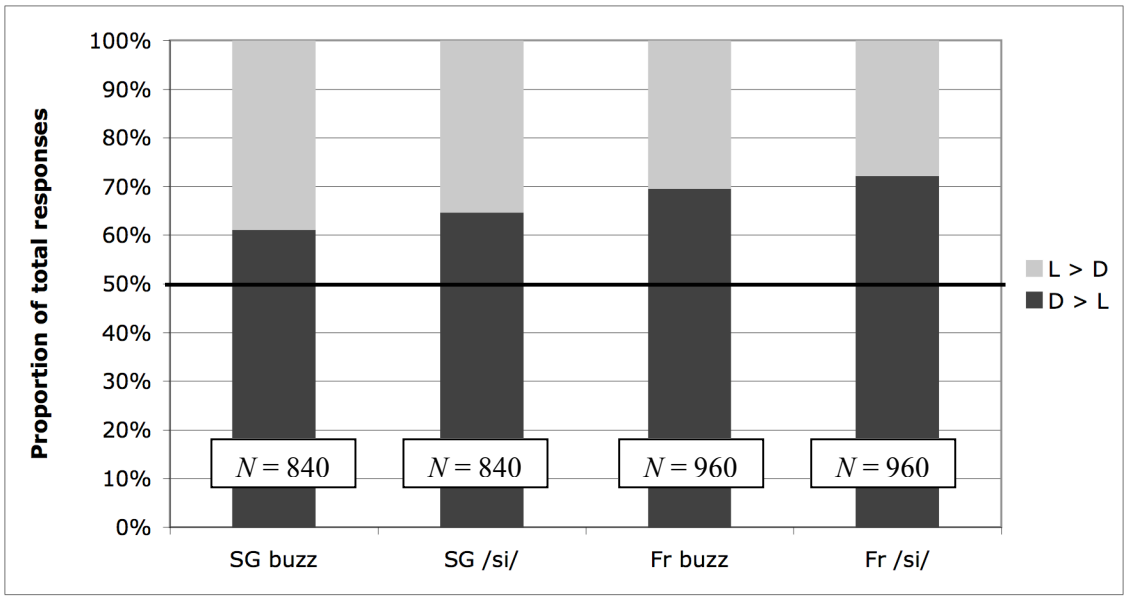

Figure 3 – Proportion of 'D > L' responses compared to the level of chance

According to binomial probabilities approximated from the standard normal distribution, subjects (in both groups for both stimuli types) responded 'D > L' significantly above the level of chance (50%) (*SG buzzes*: $z = 6.42$, $p < 0.001$; *SG /si/*: $z = 8.49$, $p < 0.001$; *Fr buzzes*: $z = 12.14$, $p < 0.001$; *Fr /si/*: $z = 13.75$, $p < 0.001$). A two-way mixed measures ANOVA with the factors *sound* (buzz, /si/) (repeated-measures) and *language* (SG, Fr) (between-groups) was conducted on the mean proportion of 'D > L' responses. No main effect of *sound* was found [$F(1,28) = 2.445$, $p > 0.05$], nor an interaction of *sound* x *language* [$F(1,28) = 0.054$, $p > 0.05$], and the effect of *language* failed to reach significance [$F(1,28) = 3.758$, $p = 0.063$].

The following explains why the two groups of Fr subjects were treated together (see §2.1). A one-way between-groups ANOVA comparing the mean proportion of 'D > L' responses for buzzes across three language groups (SG, Swiss Fr and Fr[2]) gave a non-significant main effect [$F(2,27) = 2.281$, $p > 0.05$]. Another one-way between-groups ANOVA comparing the mean proportion of 'D > L' responses for /si/ across three language groups gave a just significant main effect [$F(2,27) = 3.496$, $p = 0.05$]. Post-hoc (Tukey HSD) tests revealed that this effect came

[2] Unequal sample sizes are automatically adjusted in *SPSS*. Furthermore a Levene test showed that homogeneity of variance was not violated ($p = 0.48$).

from the difference between the SGs and Swiss Fr ($p = 0.05$), not the SGs and Fr ($p > 0.05$), nor, most importantly, the Swiss Fr and Fr ($p > 0.05$).

Out of the variables other than native language which were included in the design, the two most interesting findings concerned direction of f0 change and duration of stimuli. Firstly considering direction of f0 change, figure 4 displays the percentage of 'falling is longer than level' ('F > L'), 'rising is longer than level' ('R > L'), and 'complex is longer than level' ('C > L') responses.

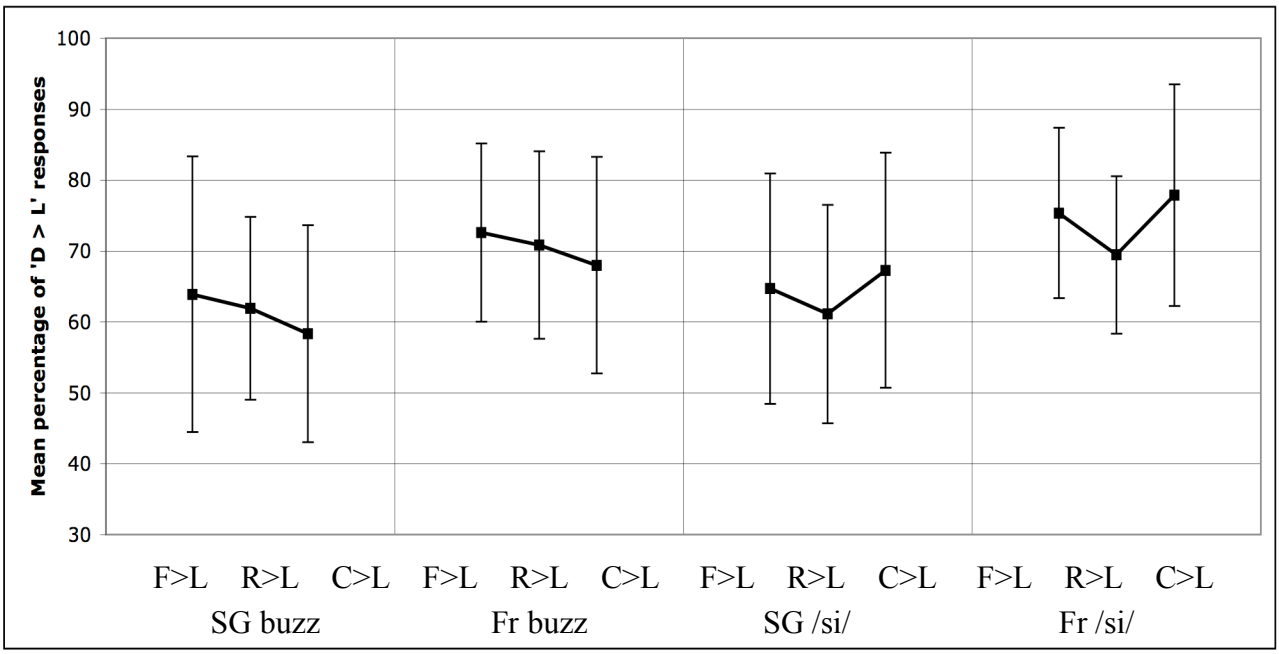

Figure 4 – Mean and sd (error bars) of 'D > L' responses depending on direction of f0 change, calculated out of total trials for each direction (F: $n = 18$, R: $n = 18$, C: $n = 24$), across 14 SG and 16 Fr subjects.

The pattern of results seems different for buzzes and /si/ monosyllables, but similar for both language groups. A three-way mixed measures ANOVA was calculated with the factors: *sound* (buzz, /si/), *direction* (fall, rise, complex) (both repeated-measures), and *language* (SG, Fr) (between-groups). Since this dataset violated the assumption of normality, logarithmically transformed values were used in the ANOVA. There were no main effects [*sound*: $F(1, 28) = 1.529$, $p > 0.05$; *direction*: $F(2,56) = 1.281$, $p > 0.05$; *language*: $F(1,28) = 3.979$, $p = 0.06$], nor interactions with *language*; however there was an interaction of *sound* x *direction* [$F(2,56) = 3.438$, $p < 0.05$]. Separate within-groups contrasts revealed that responses to falls and rises were not significantly different from one another in the two *sound* conditions [$F(1,28) = 0.701$, $p > 0.05$], however the proportion of 'C > L' responses was significantly higher than that of 'R > L' responses in the /si/ (but not the buzz) condition [$F(1,28) = 8.184$, $p < 0.01$].

Secondly considering the duration of stimuli, recall that each trial comprised two stimuli of equal duration, but for each of the different dynamic f0 contours, there were three different durations of stimulus, each paired with an equal length level stimulus. Figure 5 displays the percentage of 'dynamic is longer than level' responses at 250ms, 375ms and 500ms.

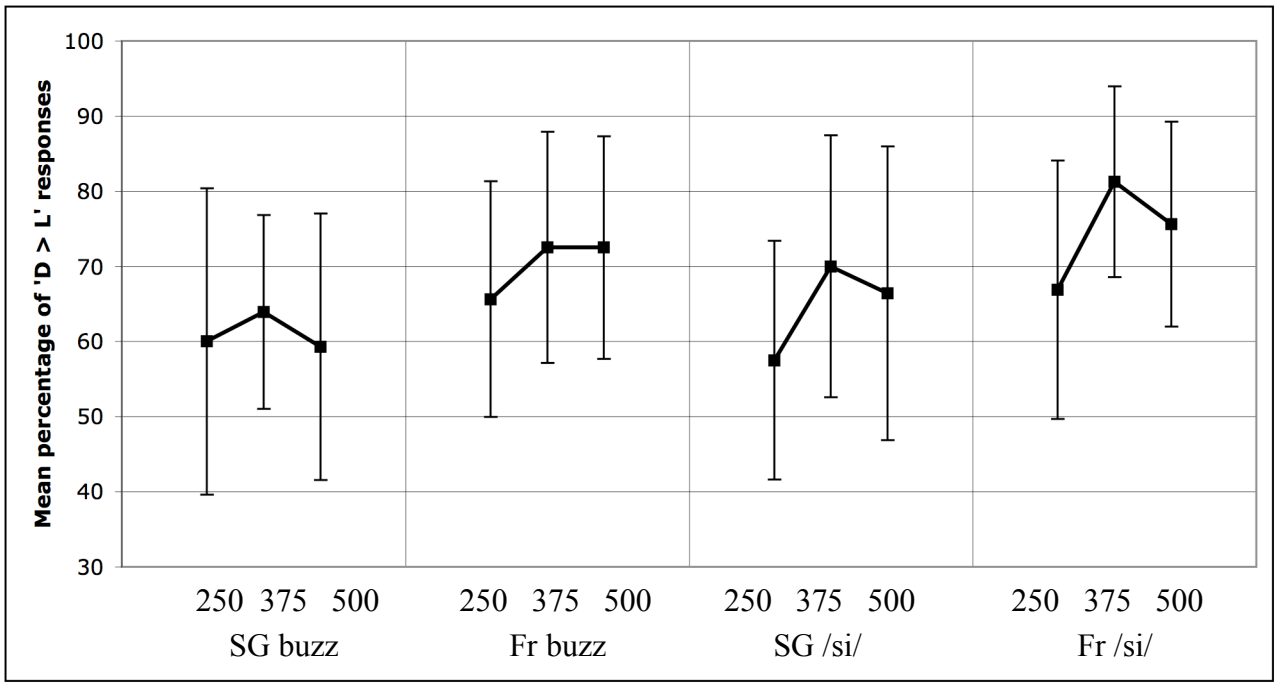

Figure 5 – Mean and sd (error bars) of 'D>L' responses depending on duration of stimuli, calculated out of total trials for each duration (250ms: $n = 20$, 375ms: $n = 20$, 500ms: $n = 20$), across 14 SG and 16 Fr subjects.

Here the pattern of results seems similar for buzzes and /si/ monosyllables, and in both language groups. A three-way mixed-measures ANOVA was calculated with the factors *sound* (buzz, /si/), *duration* (250ms, 375ms, 500ms) (both repeated-measures), and *language* (SG, Fr) (between-groups). There were no interactions, nor main effects of *sound* or *language* [*sound*: $F(1,28) = 2.425$, $p > 0.05$; *language*: $F(1,28) = 3.726$, $p = 0.06$], however there was a main effect of *duration* [$F(2,56) = 6.659$, $p < 0.01$]. Individual within-subjects contrasts for this variable revealed that the proportion of 'D > L' responses at 375ms was significantly higher than at 250ms [$F(1,28) = 11.713$, $p < 0.01$], but not quite significantly higher than at 500ms [$F(1,28) = 3.746$, $p = 0.06$].

Briefly considering timing of f0 change, four separate two-way repeated-measures ANOVAs with the factors *direction* (fall, rise) and *timing* (late, early) – one for each language group and type of stimuli – revealed some interesting interactions of *timing* x *direction,* although no main effects. For buzzes, Fr speakers judged late falls and early rises as longer than level significantly more often than late rises and early falls [$F(1,15) = 10.472$, $p < 0.01$], and SGs also showed this interaction, although not at a significant level [$F(1,13) = 1.270$, $p > 0.05$]. For /si/ monosyllables the opposite effect occurred (late rises and early falls were judged as longer than level significantly more often than late falls and early rises); this was significant for SGs [$F(1,13) = 9.244$, $p < 0.01$], but marginal for Fr speakers [$F(1,15) = 4.123$, $p = 0.06$].

There seemed to be no effect of excursion of f0 change, since a three-way mixed-measures ANOVA with the factors *sound* (buzz, /si/), *excursion* (steep, shallow) (both repeated-measures), and *language* (SG, Fr) (between-groups) revealed no main effect of excursion [$F(1,28) = 0.050$, $p > 0.05$]. There was a main effect of *sound* [$F(1,28) = 6.513$, $p < 0.05$], which corresponds to the finding for direction of f0 change: when a stimulus has a complex f0 contour, it is more likely to be judged as longer than a level stimulus in the /si/ condition compared to the buzz condition.

Since it had previously been found that order of stimuli within each trial, and order of hearing buzzes and /si/ monosyllables could influence results, these factors were tested statistically. However no effects of either type of ordering were found in these data. Finally, the results of the control trials with two identical (level) stimuli were analysed, illustrated in figure 6.

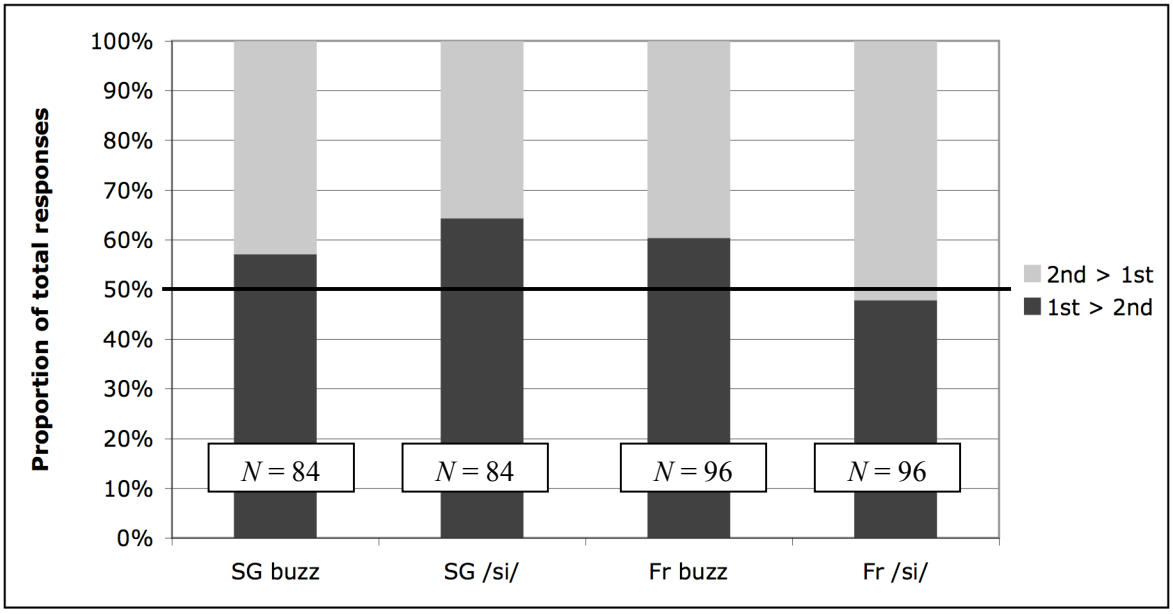

Figure 6 – '2nd is longer than 1st' compared to '1st is longer than 2nd' responses.

Over all buzz trials, SG and Fr-speaking listeners judged the first stimulus as longer than the second, above the level of chance (50%), however this was only significant for Fr [Binomial tests: *SG buzzes*: n = 84, $p > 0.05$; *Fr buzzes*: n = 96, $p < 0.05$]. Over all /si/ trials, SGs judged the first stimulus as longer than the second at a level significantly above chance [n = 84, $p < 0.01$], whereas Fr speakers judged the second stimulus as longer than the first, but not significantly above chance [n = 96, $p > 0.05$].

4 DISCUSSION

4.1 Native language

It was predicted that SGs would generally perceive dynamic stimuli as longer than equally long stimuli with level f0, whereas Fr speakers would not, and thus the two groups' responses would be significantly different. The results provide evidence in support of the SG but not the Fr hypothesis, because (i) the Fr speakers perceived more dynamic stimuli as longer than level stimuli compared to the SGs, and (ii) this difference failed to reach significance. Most importantly though, both groups perceived dynamic stimuli as longer than level stimuli significantly above chance.

This result raises two points worthy of discussion. Firstly, did listeners relate their responses to their language-specific stress perception (cf. Lehiste 1976)? If f0 movement cues 'initial' stress and duration cues 'final' stress in Fr, there would be no reason for Fr speakers to associate increased duration with f0 movement, yet these appear to be doing just that. However, there is relatively little perceptual (and

production) data to substantiate these claims about Fr stress, therefore more research would be needed before drawing conclusions on this matter.

Secondly, there is little evidence that a perceived lengthening effect of dynamic f0 is dependent on native language, since these two groups of listeners have prosodically different native languages yet both groups showed this effect. Nevertheless, many more language backgrounds would need to be tested before we could claim this was universal in speech perception, especially since this finding was not replicated with Swedish, German, Thai and Spanish listeners. Perhaps there is an underlying cause, other than vowel length contrasts (cf. Lehnert-LeHouillier 2007), language-specific stress cues, rhythm 'class' or intonational characteristics, that leads to different responses from different language groups. If this were the case, it could explain why a cross-linguistic difference did not occur in the present data. It would be interesting to investigate how speakers of a tone language, in which f0 movement cues phonological contrasts, respond to stimuli identical to those used in the present study.

4.2 Other variables

An interesting finding emerged regarding the linguistic/non-linguistic nature of the stimuli and the direction of f0 change. There was an interaction of these two factors, in that complex f0 contours were judged as longer than level significantly more often than rises were judged as longer than level in the linguistic condition compared to the non-linguistic condition. This suggests that we should be cautious in comparing the results of the previous studies which all differed methodologically in these respects: Lehiste (1976) used synthetic vowels with complex f0 contours, Pisoni (1976) and Wang *et al*. (1976) used similar stimuli but with falls or rises; Rosen (1977b) and van Dommelen (1991, 1993) used meaningful words with falls or rises; Lehnert-LeHouillier (2007) and Yu (2006) used nonsense words with falls or rises.

Furthermore, timing of f0 change appeared to have a different effect in the linguistic and non-linguistic conditions. For buzzes, late falls and early rises were judged longer then level significantly more often than late rises and early falls were judged longer than level (although this effect was only statistically significant for Fr speakers), whereas the opposite effect occurred for /si/ monosyllables. These trials were included to test for the effect identified by House (1990), that during periods of spectral change f0 movements are perceived as level, whereas during spectral stability they are perceived as changing. Logically, then, the two contours judged as longer than level most often in the /si/ condition would be perceived as follows: late rises as rising f0, since spectral stability would occur by the time the rise started; early falls as low, level f0, since spectral stability would only occur after the fall when f0 was already low. (Of course there is no spectral change in the buzzes.) However this does not correspond to the fact that in this study rises are generally judged longer than level less often than falls and complex contours. It is not clear from this data whether tonal perception during spectral change consequently affects duration perception.

Stimuli duration also affected responses significantly. In pairs of stimuli at 375ms, the dynamic one was perceived as longer than level more often than in pairs at 250ms and 500ms, particularly for /si/ monosyllables. Evidence from production data used in the creation of stimuli for another experiment (see Cumming in preparation) suggests 375ms is a natural length for /si/ uttered at a normal rate. Stimuli at 250ms probably seemed very short, and those at 500 ms very exaggerated. In both cases subjects may have responded in a more random manner than at 375ms.

5 CONCLUSION

The aim of this study was to investigate whether f0 interacts with duration in perception, and to reiterate that this has implications for duration-based rhythm research. Since length judgments do not differ significantly between the two (prosodically different) language groups, this tends to suggest that the perceived lengthening effect of dynamic f0 is not dependent on language background, although it may not be universal. Further research on SG and Fr listeners is currently investigating whether this apparent perceptual interaction between f0 and duration extends to perception at the rhythmic-group level (Cumming in preparation).

Nevertheless, if f0 changes affect the subjective duration of successive intervals in several languages, the rhythm of a language which tends to use f0 dynamism within the syllable may be perceived differently from that of one in which f0 changes minimally within the syllable. However, durational rhythm metrics may not accurately reflect this difference; therefore, in answer to the title, rhythm metrics should take account of f0. Finding a suitable means of integrating duration and f0 change into metrics such as the PVI could be the next challenge.

6 References

Adams, C. (1979) *English Speech Rhythm and the Foreign Learner.* The Hague: Mouton.

Andreassen, H. N. (2006) Aspects de la durée vocalique dans le vaudois. *Bulletin PFC: Phonologie du Français Contemporain - Usages, Variétés et Structure* 6: 115-134.

Benguerel, A.-P. (1971) Duration of French vowels in unemphatic stress. *Language and Speech* 14: 383-391.

Benguerel, A.-P. (1973) Corrélats physiologiques de l'accent en français. *Phonetica* 27: 21-35.

Bertinetto, P. M. and C. Bertini (2008) On modeling the rhythm of natural languages. *4th Conference on Speech Prosody.* Campinas, Brazil.

Cumming, R. E. (in preparation) *A Cross-linguistic Study of Rhythm: French and Swiss German*. Doctoral thesis, University of Cambridge.

Dauer, R. (1983) Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* 11: 51-62.

Dauer, R. (1987) Phonetic and phonological components of language rhythm. *11th International Congress of Phonetic Sciences.* Tallinn, Estonia.

Delattre, P. (1966) *Studies in French and comparative phonetics*. Heidelberg: Groos.

Dellwo, V. (2006) Rhythm and speech rate: A variation coefficient for deltaC. In P. Karnowski and I. Szigeti (eds.) *Language and Language Processing: Proceedings of the 38th Linguistic Colloquium.* Frankfurt: Peter Lang.

Di Cristo, A. and D. J. Hirst (1997) L'accentuation non-emphatique en français: stratégies er paramètres. In J. Perrot (ed.) *Polyphonie pour Ivar Fónagy.* Paris: L'Harmattan.

Galloway, R. E. (2007) *Bilinguals' interacting phonologies? A study of speech production in French~Swiss German bilinguals*. Unpublished MPhil thesis, University of Cambridge.

Grabe, E. and E.L. Low (2002) Durational variability in Speech and the Rhythm Class Hypothesis. In N. Warner and C. Gussenhoven (eds.) *Papers in Laboratory Phonology 7.* Berlin: Mouton de Gruyter.

Grosjean, F., S. Carrard, C. Godio and L. Grosjean (2007) Long and short vowels in Swiss French: their production and perception. *French Language Studies* 17: 1-19.

Handbook of the International Phonetic Association (1999) Cambridge: Cambridge University Press.

Häsler, K., I. Hove and B. Siebenhaar (2005) Die Prosodie des Schweizerdeutschen – Erkenntnisse aus der sprachsynthetischen Modellierung von Dialekten. *Linguistik online* 24: 187-224.

House, D. (1990) *Tonal Perception in Speech.* Lund: Lund University Press.

Lehiste, I. (1976) Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics* 4: 113-117.

Lehnert-LeHouillier, H. (2007) The influence of dynamic F0 on the perception of vowel duration: cross-linguistic evidence. *16th International Congress of Phonetic Sciences.* Saarbrücken, Germany.

Low, E. L. (1998) *Prosodic prominence in Singapore English*. Unpublished doctoral thesis, University of Cambridge.

Mertens, P. (1991) Local prominence of acoustic and psychoacoustic functions and perceived stress in French. *12th International Congress of Phonetic Sciences.* Aix-en-Provence, France.

Parmenter, C. E. and A. V. Blanc (1933) An Experimental Study of Accent in French and English. *PMLA* 48: 598-607.

Pike, K. (1945) *The Intonation of American English.* Ann Arbor: University of Michigan Press.

Pisoni, D. B. (1976) Fundamental frequency and perceived vowel duration. *Journal of the Acoustical Society of America* 59: S39.

Ramus, F., M. Nespor and J. Mehler (1999) Correlates of linguistic rhythm in the speech signal. *Cognition* 73: 263-292.

Rigault, A. (1962) Rôle de la fréquence, de l'intensité et de la durée vocalique dans la perception de l'accent en français. *4th International Congress of Phonetic Sciences.* Helsinki, Finland.

Rosen, S. M. (1977a) The Effect of fundamental frequency patterns on perceived duration. *Speech Transmission Laboratory Quarterly Progress and Status Report* 1: 17-30.

Rosen, S. M. (1977b) Fundamental frequency patterns and the long-short vowel distinction in Swedish. *Speech Transmission Laboratory Quarterly Progress and Status Report* 1: 31-37.

Schmid, S. (2001) Un nouveau fondement phonétique pour la typologie rhythmique des langues? Poster presented at the conference for the *10ème anniversaire du Laboratoire d'Analyse Informatique de la Parole (LAIP).* Lausanne.

Siebenhaar, B., M. Forst and E. Keller (2004) Prosody of Bernese and Zurich German. What the development of a dialectal speech synthesis system tells us about it. In P. Gilles and J. Peters (eds.) *Regional Variation in Intonation.* Tübingen: Niemeyer Verlag.

Vaissière, J. (1991) Rhythm, accentuation and final lengthening in French. In J. Sundberg, L. Nord and R. Carlson (eds.) *Music, Language, Speech and Brain: proceedings of an International Symposium at the Wenner-Gren Center, Stockholm, 5-8 September 1990.*

Van Dommelen, W. (1991) F0 and the perception of duration. *12th International Congress of Phonetic Sciences.* Aix-en-Provence, France.
Van Dommelen, W. (1993) Does dynamic F0 increase perceived duration? New light on an old issue. *Journal of Phonetics* 21: 367-386.
Wallin, J. E. W. (1911) Experimental Studies of Rhythm and Time. *Psychological Review* 18: 100-119.
Wang, W. S.-Y., I. Lehiste, C.-K. Chuang and N. Darnovsky (1976) Perception of vowel duration. *Journal of the Acoustical Society of America* 60: S92.
White, L. and S. Mattys (2007) Calibrating rhythm: First language and second language studies. *Journal of Phonetics* 35: 501-522.
Wiese, R. (2000) *The Phonology of German.* Oxford: Oxford University Press.
Woodrow, H. (1911) The Role of Pitch in Rhythm. *Psychological Review* 18: 54-77.
Yu, A. C. L. (2006) Tonal effects on perceived vowel duration. *Laboratory Phonology 10.* Paris, France.

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

reg50@cam.ac.uk

http://www.ruth.galloway.me.uk/Site/Research.html