Chapter 5

Real-World Data Analysis: A Case Study

Introduction

The discussion in the previous chapters have shown that there are:

- many situations and conditions in which tempo variation can be observed, e.g. text sort or emotive speech
- many mechanisms with which tempo variation can be achieved, e.g. by varying pauses or reducing the duration of vowels and consonants
- methodological problems in measuring tempo, e.g. how to indicate tempo values for various inter-pause stretches for spontaneous speech.

In this chapter a case study with real-world data is given, namely a prosodic analysis of horse race commentaries. The advantages of this somewhat special text sort are the following ones:

- perceptually, the impression is of continously increasing tempo
- it is emotive speech, and the integration of emotions in synthetic speech is one type of application of tempo-scaling in synthetic speech
- it is real-world data, not laboratory speech or a meta-form of speech

The intriguing question is how the tempo increase that iscontinously manifest over the entire commentary is achieved. Moreover, other prosodic effects can also be observed, not just those realised on a durational level. Thus, the question can be extended to "how do prosodic systems interact?"

5.1. The prosody of excitement in horse race commentaries

As mentioned in chapter 2, high emotional activity in speech is characterised by a higher pitch average, a wider pitch range, a higher intensity, and a faster speech rate compared to some neutral or default way of speaking (Banse & Scherer, 1996; van Beezooyen, 1984; Murray & Arnott, 1993). Terms used to describe excitement or arousal are e.g. "anger", (especially "hot anger" or "rage") on the negative side, and "elation", "joy" and "happiness" on the positive side.

A further form of arousal, which is investigated in this study, could be named "suspense" and is exemplified in horseracing commentaries (henceforth HRC).

Prosody of horse race commentaries

In HRC, the course of emotional arousal is dependent on the fixed framework of the race, progressing from relative calm, through increasing excitement to the climax at the finish, then returning to a post-race calm. The form of arousal is presumably specific to sports commentaries, being neither negative (as with "fear" or "hot anger") nor positive (as with "elation"), but expressing the commentator's sense of excitement and suspense.

Barry (1995) gives an auditorily based description of a typical HRC pattern:

> "In British English there's a clear monotonisation rule, with definite, race-stage oriented resets (to a high pitch) with tempo and volume increases from one series of 'intonation units' to the next, and with sudden rallentando and decrescendo combined with a short series of resets to a lower pitch and a final low falling contour from the moment the winning horse finishes."

A stylised HRC pattern features an increase of pitch level, tempo and intensity, and a decrease of pitch range (figure 5.1).

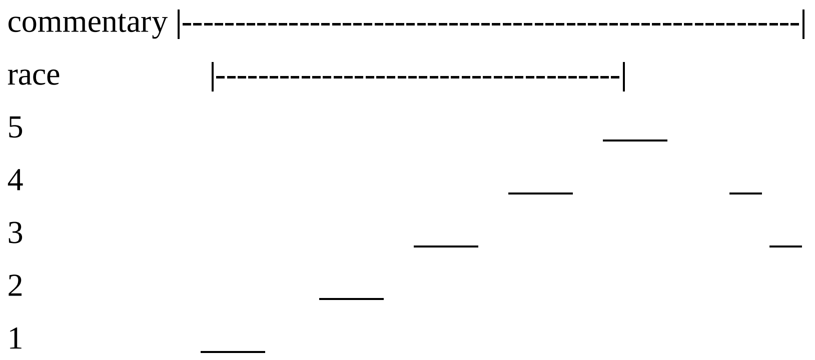

Figure 5.1: Stylised time course of horse race commentary according to Barry (1995). Parameters: pitch level (1 = normal, 5 = high); pitch range (1 = normal, 5 = narrow); tempo (1 = normal; 5 = very fast); intensity (1 = normal; 5 = high).

Spontaneity of horse race commentaries

The advantage of sports commentaries is that they offer examples of non-scripted speech with genuine informational and emotional expression. Although one cannot claim that HRC is completely "spontaneous" since it has a conventionalised form, it is taken from real situations compared to emotive speech acquired in laboratory situations from actors, which is the usual way of eliciting emotive speech production.

A further advantage of sports commentaries is the quality of the speaker, which guarantees a certain level of fluency, allowing an analysis of prosodic structuring which is less disturbed by dysfluencies (e.g. filled pauses, false starts, mispronunciations, syllable drawls, repeats, ungrammatically located pauses) than is the case with non-professional speakers.

Purpose of this study

The present study pursues a number of aims: Firstly, it seeks a verification of the auditory description of HRC given in Barry (1995). Secondly, it offers quantitative data on a number of prosodic parameters, among them several tempo-related parameters, which vary systematically with degree of excitement/arousal within a situation. This provides independent evidence for the pattern of increasing and subsiding excitement reflected by those parameters. Thirdly, it thus provides an empirical basis for comparison with other forms of excitement.

We therefore attempt to identify and discuss (1) what is specific to the HRC speaking style, (2) what is specific to the excitement component of the HRC.

5.2. Methods

Material

Three horse-race commentaries were chosen at random from a selection of HRCs recorded during BBC Grandstand (television) transmissions. The durations (time period from start of the race to the moment the winning horse finishes) of the three HRCs range from 79 seconds for race 1 to 145 seconds for race 3. Each race was commentated by a different (male) speaker. Although their identities are not known to the author, the speakers in race 2 and 3 are recognisable as speakers of New Zealand and Australian English, respectively. The commentator in race 1 speaks a standard southern English accent.

Analysis

In the first analysis step, three phases were selected to encompass the build-up and decline of tension expressed in the commentary. These are located

1) near the beginning of the race
2) in the middle
3) spanning the finish until the moment the winning horse passes the line.

These phases are sandwiched between breath pauses and they comprise around 25 seconds in the first race, 26 seconds in race 2, and 32 seconds in race 3.

For each phase an orthographic transliteration with special symbols for silent pauses, breath pauses, and filled pauses such as [ɛː] was carried out. Filled pauses were regarded as articulation, so that only silent and breath pauses are considered as ("unfilled") pauses below.

Each pause was marked in the time course, providing data on the number of pauses, the pause durations, and the durations of inter-pause stretches and inter-breath stretches, respectively.

In order to determine speech rate, the number of phonological syllables for each inter-pause stretch was counted. Filled pauses were regarded as one articulated syllable. Thus, "speaking rate" is defined as the number of syllables divided by total speaking time including pauses - in contrast to "articulation rate", which is based on total speaking time excluding pauses.

Prosodic phrases demarcated by intonational and rhythmical means other than pauses were also transcribed. However, these phrase markers were used for further instrumental investigation due to the lack of reliable labelling criteria.

In the second analysis step, instances of prominent horse names were excised. For race 1 "Two Clubs" occurred six times, in race 2 "Pentland's Flyer" 11 times, and race 3 "Hot 'n' Saucy" 15 times. These occurrences, which are more or less equally distributed over the entire race, were used to determine the pattern of pitch and intensity characteristics.

F0 range, defined as the difference between the highest and lowest value in comparable syllables, was measured in the second stressed syllable of each horse name.

Spectral tilt as an acoustic correlate of glottal excitation, was calculated as the difference in dB between the amplitude of the first harmonic (H1) and the amplitude of the second formant (A2) (cf. van Sluijter & van Heuven, 1996). Only [a]-like sounds were considered: [ʌ] in "Two Clubs", the [a]-portion of the diphthong in "Pentland's Flyer", and [ɑ] in "Hot 'n' Saucy" (Australian accent).

The same vocalic portions were used to determine intensity in dB.

5.3. Results

Filled pauses

One of the most typical phenomena of spontaneous speech is the filled pause. For all three speakers filled pauses were observed, though with differences in the number of instances (see table 5.1). Apparently, commentator 1 has a lower level of fluency compared to the commentator of race 3. The fact that all speakers show some form of dysfluency (despite their being professional speakers) can be seen as an indicator of spontaneous communication, and therefore as evidence for the naturalness of the speaking situation.

Table 5.1: Number of filled pauses and their locations with reference to the inter-pause stretch (ips). Fillers within ips ("ips-mid") usually occurred at syntactic phrase boundaries.

speaker	ips-onset	ips-mid	ips-offset	total
1	6	5	-	= 11
2	2	3	1	= 6
3	1	1	-	= 2

Duration of pauses and phrases

In general, the data (table 5.2) show that, as the race progresses, the average pause length does not get shorter in all cases. However, the inter-pause stretches do get shorter, and a parallel reduction in the inter-breath stretches reflects the increased breath rate towards the end. The corollary of this trend is, of course, more pauses per time unit as the race progresses.

Exceptions to the trend are, however, the central portion of race 1, which has longer pauses than in the beginning and end phases, and of race 2, which has both longer pauses and longer inter-pause stretches.

From table 5.2 it is evident that the shortest pause durations occur in the final part. This is in agreement with the expected tempo increase towards the end, because one strategy for achieving a higher tempo is to shorten the pauses. From the overall HRC pattern we might propose a general rule like "the later in the race, the shorter the pauses". This idea, however, is spoiled by the pause durations in the middle part of race 1 and 2 which are longer than those for the initial part.

Table 5.2: Average duration in seconds of unfilled pauses, inter-pause stretches (articulation phase between two unfilled pauses), inter-breath stretches (articulation phase between two breath pauses). Phases of approx. 25-32 seconds at the beginning, middle, and end of the race were investigated.

	race	**beg**	**mid**	**end**
pause	1	0.348	0.408	0.279
	2	0.356	0.396	0.226
	3	0.635	0.495	0.263
inter-pause stretch	1	4.303	2.349	2.063
	2	2.587	4.115	2.231
	3	3.763	3.127	2.794
inter-breath stretch	1	4.303	3.523	3.438
	2	3.880	6.173	3.069
	3	4.390	4.020	3.841

Another way of speeding up is to reduce the number of pauses, which results in longer inter-pause stretches. The expectation of "the later in the race, the fewer the pauses, the longer the inter-pause stretches" is supported by no speaker, except in race 2, again in the middle part.

The shortened inter-breath stretches in the final phase of each race points to an increase in air-flow during speech towards the end.

Global tempo

Neither for speaking rate nor for articulation rate can the expected pattern of speeding up during the commentary be confirmed (see table 5.3). For the three parts of each of the race commentaries the middle part deviates from the other two, either as the slowest phase (race 1), or as fastest phase (race 2 and 3). Even if we compare only the beginning and the final parts, only two out of six measurements show a slight increase in the syllabic rate. Obviously the global tempo measured in syllables per second does not reflect the pattern described in Barry (1995). Other factors must be responsible for the impression of gradual acceleration.

Table 5.3: Speaking rate (SR, including pauses) in syllables per second for three different phrases. Articulation rate (AR, excluding breath and silent pauses) in syllables per second for three different phrases.

	race 1		**race 2**		**race 3**	
part	**SR**	**AR**	**SR**	**AR**	**SR**	**AR**
beg	4.61	4.92	4.28	4.81	3.91	4.48
mid	3.98	4.59	4.54	4.90	4.08	4.66
end	4.58	5.14	4.10	4.48	4.08	4.43

Fundamental frequency level

Figure 5.2 demonstrates the successive build-up of tension reflected in the fundamental frequency in all three races investigated.

The most extreme example in race 2 reveals a difference of 15.3 semitones between the lowest token at the beginning and the highest token at the finish, i.e. one and a quarter octaves. Almost one octave difference (11.5 semitones) can be observed for race 3, and over half an octave (7.5 semitones) for race 1.

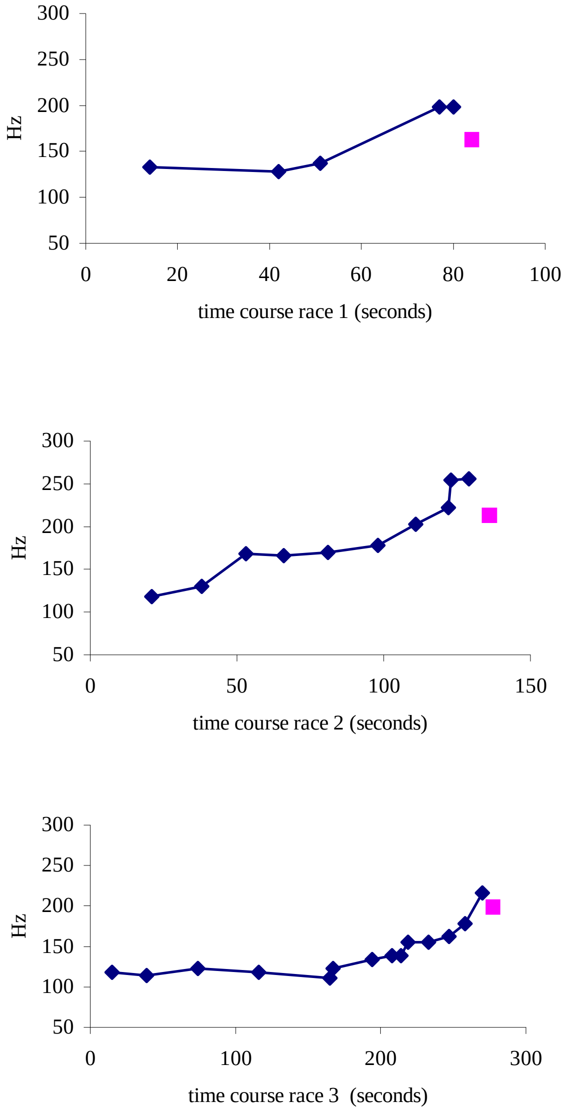

Figure 5.2: The F0 course (in Hz) at the occurrences of same words (horse name) for the three races. Values taken at the second stressed syllable are marked on the time scale (x-axis). In each graph the rightmost value is the one taken directly after the finish.

Fundamental frequency range

The F0 range, i.e. the difference between the highest and the lowest F0 value found in the examined vowels, is expected to be narrowed. We see here a variation as a function of speakers, position in the race, position in the intonational phrase, and length of the investigated vowel. Speaker 1 gradually reduces the range from 9 to 0 Hz in the five examples of [ʌ] in "Clubs". Speaker 3 also shows a gradual reduction of the F0 range from 25 Hz to 5 Hz in a phonemically long [ɔː] in "Saucy". However, in the middle part small ranges (< 10 Hz) alternate with larger ranges (> 10 Hz). This pattern of alternation in the middle phase is similar for speaker 2, who shows more extreme ranges (up to 57 Hz) in the phonemically long diphthong [αI], or [αI↔], in "Flyer". Additionally, speaker 2 widens the range again towards the very end.

As far as it is possible to interpret these numbers, one can say that F0 range is not homogeneous across the race nor across the speakers. In this sense the narrowing of the F0 range described by Barry (1995) can be confirmed for speaker 1, but only partially confirmed for speakers 2 and 3.

Fundamental frequency movements

Regarding F0 movements on "nuclear" tone occurrences of the horses' names, race 1 maintains a fall in the first two occurrences, but shows quite a level tone in the remainder. This pattern of starting with a fall, and ending with a level tone is repeated in race 3. Race 2, however, starts with a fall on "Flyer" and changes to a rise pattern or, alternatively, a rise-fall pattern for the finishing part. We see here, that the default falling contour (a high-low pitch accent) alternates either with a level tone (monotonal pitch accent) or a rising contour (a low-high tone). It is debatable whether this alternation of pitch accent patterns is a change in the phonological structure (choice of tonal accent with a linguistic function) or a phonetic realisational phenomenon.

Overall intensity

Since F0 and Sound Pressure Level (SPL) are known to correlate with sub-glottal pressure (Ladefoged, 1967), an increase in intensity may be expected to accompany the observed rise in F0 over the course of the HRC. The auditory observation of increased loudness is of course a further reason to expect this. Table 5.4 confirms these expectations, if we ignore one outlier. All three speakers have in common that the intensity is lowered after the finish. This can be seen as a reduction of sub-glottal pressure, an indication of relaxing tension.

Table 5.4: Overall intensity and spectral tilt in dB at occurrences of same words (horse name) for the three races. Mean values from the first third (beg), second third (mid), last third (end) of each race are given. Numbers of occurrences (#) for each phase are indicated in the first rows. The value taken directly before, and the one directly after the finish are listed separately.

	race	**beg**	**mid**	**end**	**last before**	**first after**
no. of occur-rences	1	1	2	2		
	2	3	3	4		
	3	3	3	8		
overall intensity	1	75.6	74.9	76.4	77.1	76.7
	2	73.5	75.7	76.5	76.9	76.0
	3	72.7	74.0	75.0	74.8	73.7
spectral tilt	1	+9.8	+6.1	+0.5	-2.6	+1.9
	2	-0.3	-10.1	-13.0	-16.0	-22.0
	3	-9.6	-10.7	-11.7s	-14.6	-9.6

Spectral tilt

If we think of a gradual inclination of tension during the race, this tension should also be transmitted to the vocal folds. As a consequence, the excitation of the glottal pulse should be strengthened. Decreased spectral tilt is accepted as a correlate of the increased effort associated with accented vs. unaccented syllables (van Sluijter & van Heuven, 1996). Since the increased effort apparent in the raised F0 and increased intensity during the HRC can be seen as the result of a similar - though longer term – physiological adjustment of e.g. sub-glottal pressure, a decrease in spectral tilt can be expected as the race progresses.

In table 5.4 the tendency for the spectral tilt to decrease is clearly visible for all speakers. For speakers 1 and 3 the increasing tilt directly after the finish can be interpreted as a relaxation of tension. In contrast, speaker 2, however, seems to remain and even to strengthen this tension.

5.3. Discussion of the analysis

HRC characteristics

The general pattern of increasing excitement during the course of a horse race, with the climax at the finish, followed by a rapid relaxation after the finish is clearly reflected in the following prosodic properties (compare table 5.5): more pauses, shorter pauses, higher breathing rate, much higher F0, higher intensity, flatter spectrum.

The direct dependency of these parameter trends on the race development (and the development of tension) is clearly reflected in the occasional abrupt up-step of F0 and other values when there is a sudden and unexpected event (e.g. race 2, 3rd and 9th mention of "Pentland's Flyer").

However, despite the overall similarity of pattern in F0 and effort (intensity level), the individual levels and ranges remain a characteristic of the individual commentator.

Prosody of excitement

In some respects a HRC shares the features of other forms of excitement. Table 5.5 lists the general trends of general emotional classes which can be subsumed under the label "excitement". The values in the table are taken from similar tables in Banse & Scherer (1996), van Beezoyen (1984), Murray & Arnott (1993). There is clear agreement among the studies regarding the general tendencies of the prosodic realisations.

The differences in the temporal parameters must be seen against the length of the investigated speech samples. It seems to be a popular paradigm in emotional speech research to study one-sentence utterances. This explains why pausing behaviour is not taken into account in the tables in Cowan (1936), van Beezooyen (1984) and Banse & Scherer (1996). Cowan (1936), however, remarks that fewer and shorter pauses occur in emotional speech, whereas we found *more* shorter pauses. This difference in results show that pausing strategies, and especially breathing, might be aspects worth studying in an emotional context, as well as tempo.

Table 5.5: Changes in prosodic parameters compared to neutral speech for different forms of excited speech: last phase of horse race commentaries (HRC), based on this study; anger, joy, surprise, fear, summarised from similar tables in Banse & Scherer (1996), van Beezoyen (1984), Murray & Arnott (1993). '+' = increase; '++' = large increase; '-' = decrease; '±' = unclear; empty cell = unknown.

prosodic parameter	**HRC**	**Anger**	**Joy**	**Surprise**	**Fear**
pausing rate	+				
breathing rate	+				
pause duration	-				
tempo	±	+	±	±	+
F0 level	++	++	+	+	+
F0 range	±	++	++	+	±
intensity	+	+	+	+	±
spectral tilt	-				

Unanimous agreement exists in terms of the increased average F0, whereas the F0 range seems greatly increased for anger and joy, less increased for surprise, but unclear for fear and HRC.

Probably due to the increased F0 level one can find the increase in overall intensity in all classes of emotions, except for the unclear case of fear.

In the study of Banse & Scherer (1996), a number of other parameters relatable to spectral tilt were investigated, but the results are not directly interpretable in the framework of the present study. Banse & Scherer (1996) show that the relative amount of high to low frequency energy varies with the expression of *different* emotions. However, they do not address the question of the degree to which a single emotion category is reflected in the strength of the parameter.

In the present study we can maintain that the activity level is clearly reflected in the prosodic patterns discussed. However, as Murray & Arnott (1993) noted, active forms of emotional speech show strong tendencies to be confused with others. Thus, it is no surprise that HRC fits well with the prosodic characteristics of high activity level.

It must be conceded that a differentiated view on an emotion such as "anger", which can be further subdivided in hot anger or rage, cold anger, threat, frustration or further nuances, would show differences in the extent of the use of prosodic parameters.

Learned vs. natural excitement

HRC as a product of an individually developed, public oriented, professional speaking style leads to the important question: although the data are from the "real" world, is the excitement in a HRC "real" in the sense of spontaneous and natural?

Although HRC recordings are considered as "spontaneous" speech, because it is unscripted, there is definitely a certain degree of routine, and therefore a certain lack of spontaneity:

- As shown above, the frame and the period of "getting prosodically excited" is given.
- The presumed expectations of the listeners are present. A non-excited commentator would probably be classified as "boring" and hence "non-professional".
- The usual acceptance of being excited is shifted: the commentator is allowed to show a degree of excitement which would be seen as exaggerated in a "default" situation.

The higher breathing rate indicates a higher level of physiological activity. However, a higher breathing rate can be explained by an increased phonatory and pulmonic effort which is reflected in the higher values for spectral tilt and intensity, respectively. It remains speculation whether this special "shouting"-style is a consciously controlled and/or a trained behaviour, or whether it represents a natural verbal manifestation of the excitement which can be observed in other people towards the finish of the race as well.

Multiple functions of prosody

It is a truism that prosody has functions on various levels of speech communication: linguistically, paralinguistically, and extra-linguistically.

We are still a long way from being able to exploit this knowledge. In the area of speech synthesis, attention has been shifted from "intelligibility" to "naturalness". In the context of emotions this means e.g. overcoming the boredom of artificial voices (Trouvain, 2000), or an explicit modelling of certain emotion types (e.g. Cahn, 1989; Murray, Arnott & Rohwer, 1996; Schröder, 1999).

In our opinion, it is important to have a theoretical framework such as the linguistic - paralinguistic - extralinguistic distinction (Crystal, 1969), and to fill it with data-supported life, in order to identify what kind of information a speaker transfers to the listener by prosodic means. An example from the HRCs shows the semantic/pragmatic content located at the paralinguistic level. A HRC tends to have a rather low sound segment or word based information value, but a relatively high prosody based information value. E.g. we assume that the placement of a horse (as it appears in the example in section 3 can be (partially) decoded by prosody, if the listeners know the prosodic context of the commentary before the climax. A further indication supporting the prime importance of prosody is the fact that some words are not comprehensible (even to phonetically trained native listeners).

We conclude that a more detailed investigation in prosody can help to model specific speaking styles such as HRC, and to model specific emotions such as excitement and its derivatives.

Summary and discussion of chapter 5

Tempo variation in horse race commentaries is clearly reflected in changes of pausing behaviour. This was expected on the basis of the general statement that tempo variation is primarily variation in pausing (Goldman Eisler, 1968). However, this case study has provided evidence against the generalisation that speeding up is characterised by fewer and shorter pauses. There are more pauses in the auditorily faster last bit of those commentaries, and an important influence of breathing and average pitch level. This study allows a differentiated view on tempo stressing the role of breathing, a parameter not frequently taken into account in emotional speech research.

In view of the sub-title of this thesis, the following question is allowed: what is the use of an analysis of a special emotive and spontaneous speaking style for *text*-to-speech synthesis? First, horse race commentaries are generated with speech synthesis in some cultures, e.g. in Japan (Campbell, personal communication). Second, the results of this case study confirm the importance of pauses for analysing tempo in particular, and prosody in general. This means for speech synthesis that tempo modelling is first of all pause modelling. Third, one of the future applications of synthetic speech is to model emotive speech. Here, tempo and pausing are essential prosodic parameters along the activation axis.