Analysis of Voice Fundamental Frequency Contours of Continuing and Terminating Prosodic Phrases in Four Swiss German Dialects

Adrian Leemann 1, Keikichi Hirose 2, Hiroya Fujisaki 3

1 Hirose and Minematsu Lab, The University of Tokyo, Japan and Universität Bern, Switzerland
2 Hirose and Minematsu Lab, The University of Tokyo, Japan
3 The University of Tokyo, Japan

{leemann, hirose}@gavo.t.u-tokyo.ac.jp, fujisaki@alum.mit.edu

Abstract

In the present study, the F0 contours of continuing and terminating prosodic phrases of 4 Swiss German dialects are analyzed by means of the command-response model. In every model parameter, the two prosodic phrase types show significant differences: continuing prosodic phrases indicate higher phrase command magnitude and shorter durations. Locally, they demonstrate more distinct accent command amplitudes as well as durations. In addition, continuing prosodic phrases have later rises relative to segment onset than terminating prosodic phrases. In the same context, fine phonetic differences between the dialects are highlighted.

Index Terms: Command-response model, Swiss German dialects, talk-in-interaction, spontaneous speech

1. Introduction

Fujisaki [1] distinguishes between linguistic, paralinguistic, and nonlinguistic information carried by speech. Linguistic information is concerned with lexical, syntactic and semantic contents of the message and is discrete in nature. Paralinguistic information is concerned with the intention and attitude of the speaker and is used to modify or supplement linguistic information, while nonlinguistic information is concerned with physical and mental states of the speaker over which the speaker usually has no direct control. The information on conversational structuring such as turn-taking is thus paralinguistic, and is primarily expressed by intonation. This function of intonation has provided fertile grounds for research in the branch of conversation analysis, but the results have so far been mainly qualitative.

The present study examines features of intonation contours quantitatively within the scope of day-to-day communication. Two concepts which are central are the terms phrase and turn. Phrase is a term in syntax, but is used also in prosody. In the present paper, a phrase is meant to be a ‘prosodic phrase’ defined by Fujisaki [2] on the basis of phrase components of the F0 contour, as obtained by its analysis using the command-response model. On the other hand, we follow Goodwin [3] in defining a turn to be “the talk of one party bounded by the talk of others, […], turn-taking being the process through which the party doing the talk of the moment is changed.”

The F0 contours of prosodic phrases bear critical conversation-structuring weight [4]. Among other features, the F0 contours of prosodic phrases signal whether the speaker intends to keep the turn (continuing phrases, C-phrases), or whether he/she intends to hand over the turn (terminating, T-phrases). T-phrases are generally marked with a falling intonation [5] and occur in turn-final position, while C-phrases tend to show rises or rise-falls in phrase-final positions [6] and normally occur in turn-initial or in turn-medial positions.

In this study, we intend to find out how the parameters of F0 contours of C- and T-phrases differ quantitatively in a given corpus, by their analysis using the command-response model [7], and at the same time, we want to show how four dialects of Swiss German differ from each other in marking C- and T-phrases.

2. Methods

Forty-one speakers from 4 dialect regions in Switzerland (10 speakers per dialect) - 2 Alpine varieties, Valais (VS) and Grisons (GR), and 2 Midland dialects, Bern (BE) and Zurich (ZH) - were recorded in spontaneous interviews where the interviewer asked questions such as “What do you plan to do after the Baccalaureate?”, “What will you do in your next vacation?” etc. The recordings were subsequently transcribed and segmented in Praat [8]. The recordings were annotated with variables that are assumed to occupy a crucial role for a cross-dialectal intonational description – one of the variables being prosodic phrase type: C- and T-phrases. The recordings, approximately 3 minutes of spontaneous speech per speaker (41,000 syllables total) was parametrized by means of Mixdorff’s FujiParaEditor [9]. The approach for the labeling of the prosodic phrases will be explained at 2.1, the command-response model will be introduced at 2.2.

2.1. Categorization of C- and T-phrases

The categorization of C- and T-phrases in the present study is performed according to syntactic and turn-related criteria and is based on [4]. For a prosodic phrase to be labeled T it has to be syntactically complete as well as suggest finality in the context of a given turn. While syntactically complete entails that the structural syntactic positions are occupied, the criteria for turn-finality are more difficult to establish. In the present study, turn-finality entails that the conversational context must clearly imply an end of a conversational activity, e.g. an end of a narration or a sub-topic. A turn-take can, but need not, occur after a prosodic phrase which fulfills the described syntactic and turn-related criteria. Figure 1 shows the criteria according to which prosodic phrases were labeled as T or C.
2.2. The command-response model

The command-response model [7] is hierarchically structured and formulated as a linear model, consisting of two second-order critically damped filters. As input signals, the model receives phrase commands (PCs) in the form of impulse functions and accent commands (ACs) in the form of rectangular functions. The input signals are processed by the phrase and accent control mechanisms. The output signals of the two mechanisms are added onto the smallest asymptotic value ($F_b$) of the $F_0$ contour that is to be generated. For analysis purposes, the model decomposes the $F_0$ contour into a set of components from which timing and frequency information of the $F_0$ contours can be estimated.

In the present study, $F_b$ is assumed to be a speaker-specific parameter, $\alpha$, the natural angular frequency of the phrase control mechanism, is set at 2/sec, $\beta$, the natural angular frequency of the accent control mechanism, at 20/sec. The phrase component can be applied for a description of the global declination tendency of $F_0$. The unit in which declination is observed in the present approach is the prosodic phrase. The accent component is understood as a device for marking segments more $F_0$-prominent on the local level.

A few critical qualifications as to the linking of ACs with the segments need to be made. Despite the fact that the literature on the acoustic properties of stress, particularly the significance of $F_0$, suggests no clear-cut relationship between stress and pitch movements (cf. [10]) it is nevertheless commonly assumed that stress is primarily a function of $F_0$ [10]. Consequently, in a framework of intonation studies applying the command-response model, ACs are commonly anchored with stressed syllables. In the present study, word stress and $F_0$ movements are treated as independent variables. As a result, ACs are not anchored with stressed syllables, instead, they are linked with exactly those syllables which they span, i.e. those syllables which are made more $F_0$-prominent.

This approach was chosen because, as shown above, the role of $F_0$ in stress-marking is not fully understood. Secondly, a study on the acoustic properties of different Swiss German dialects’ realizations of word stress, which would possibly justify an anchoring of ACs on the level of stressed syllables if the primary acoustic correlate of stressed syllables were indeed $F_0$, has not yet been conducted. Finally, earlier studies on Swiss German (cf. [11]) have shown significant incongruencies between the placement of word stress and pitch movements.

3. Results

The two PC parameters examined in the present study are PC magnitude and prosodic phrase length. The temporal positioning of the PC will not be further addressed, because its position, at least in utterance-initial position, is fixed at 500 msec before segment onset (since $T_0 = 1/\alpha$ if, as it is the case in the present study, the peak of the phrase component is positioned at prosodic phrase-onset and $\alpha$ is held constant). For ACs, the timing parameters $T_1$ and AC duration as well as AC movement and AC amplitude are examined, since they reflect the lexical information of the words.

In order to obtain more specific results, the ACs were labeled according to their position in the prosodic phrase: first, medial, penultimate, and ultimate. The results presented here are from prosodic phrases which included 2 ACs or more. In other words, the following results are based on analyses which do not take the number of ACs contained in the prosodic phrase into account. Note that AC amplitude and PC magnitude values were square-root transformed to fulfill the requirements of parametric tests.

3.1. Phrase commands

3.1.1. Magnitude

Compared to T- phrases, C- phrases overall show significantly higher magnitudes ($t$-test $t(4790) = -9$, $p < .0001$). Figure 3 shows the ANOVAs on the dialects’ PC magnitudes, according to prosodic phrase type. The vertical endpoints of the diamonds indicate the 95% confidence intervals for each group’s mean. The horizontal line in the figure represents the total response sample mean. Both ANOVAs confirmed significant effects (T-phrases: $F(3, 1576) = 26.5$, $p < .0001$, C- phrases: $F(3, 3208) = 26.8$, $p < .0001$).

In T-phrases, the GR demonstrate the highest declination (+10% from the total response mean), while the ZH show the lowest magnitudes (-6%). The same trend is found in the continuing condition.
3.1.2. Duration

Compared to T-phrases, C-phrases demonstrate significantly shorter durations (sec) (t-test \( t(4790) = 3.8, p < .0001 \)). Figure 4 indicates the ANOVAs on the dialects’ PC durations, according to prosodic phrase type. Both ANOVAs showed significant effects (T-phrases: \( F(3, 1578) = 3.4, p = .02 \), C-phrases: \( F(3, 3210) = 5.8, p = .0066 \)).

Figure 4: ANOVA on the dialects’ PC durations according to prosodic phrase type.

In T-phrases, the VS exhibit the longest durations (+5%) while the GR demonstrate the shortest IPs (-9%). In C-phrases, the ZH group indicates the longest IPs (+8%), while the VS show the shortest IPs (-5%).

3.2. Accent commands

3.2.1. Amplitude

Compared to T-phrases, C-phrases demonstrate significantly higher AC amplitudes (t-test \( t(11145) = -8, p < .0001 \)). Figure 8 shows the ANOVAs on the different AC positions’ amplitudes. Both ANOVAs confirm significant effects (T-phrases: \( F(3, 3288) = 50.6, p < .0001 \), C-phrases: \( F(3, 6526) = 284, p < .0001 \)).

Figure 5: ANOVAs on the different AC positions’ amplitudes, according to C-phrases (left) and T-phrases (right).

Figure 5 shows that in T-phrases, we find a rise-fall-rise pattern in AC amplitude, while the overall pattern in C-phrases is rise-rise-rise in AC amplitude. In both, C- and T-phrases, the first accent command is low in amplitude, followed by a significantly higher medial AC in both instances – in the case of C-phrases the increase is 24% (AC medial position compared to AC first position), while in the T-condition, the increase is only 15%. The pattern that follows after the medial AC is different for both prosodic phrase types. If the same calculations are made for each dialect, we find an increase-increase-increase pattern for the GR and VS dialects in T-phrases, the BE demonstrate a decrease-increase-increase pattern and the ZH an increase-increase-decrease pattern. In C-phrases, all dialects show the usual increasing pattern, only the GR group demonstrates an increase-decrease-increase pattern.

3.2.2. Duration

Compared to T-phrases, C-phrases show significantly longer AC durations (t-test \( t(11145) = -5.5, p < .0001 \)). Figure 6 gives the ANOVAs on the AC positions’ durations. Both ANOVAs prove significant effects (T-phrases: \( F(3, 3288) = 16.4, p < .0001 \), C-phrases: \( F(3, 6526) = 71, p < .0001 \)).

Figure 6: ANOVAs on the different AC positions’ amplitudes, according prosodic phrase type.

Figure 6 shows that in both, continuing and terminating prosodic phrases, we find an overall boost of AC duration from AC position to AC position, while the increases are more distinct in the continuing IPs. If the same calculations are made for each dialect, we find an increase-increase-increase pattern for the GR and VS dialects in T-phrases, the BE demonstrate a decrease-increase-increase pattern and the ZH an increase-increase-decrease pattern. In C-phrases, all dialects show the usual increasing pattern, only the GR group demonstrates an increase-decrease-increase pattern.

3.2.3. Movement

Overall, the vast majority of ACs are rising (87.4%). A two-tailed Fisher’s exact test confirmed significant differences in the relative proportions of rising and falling ACs between C- and T-phrases (p < .0001), C-phrases illustrate 12%, T-phrases 16% falling ACs. In all dialects, T-phrases demonstrate more falling ACs than continuing prosodic phrases (not sig. for ZH). More detailed analyses are shown in Figure 7, where the mosaic plots for AC position and AC movement type are shown according to prosodic phrase type.

Figure 7: Mosaic plot of AC position and AC movement type by prosodic phrase type.

Figure 7 shows that in T-phrases, we find a rise-fall-rise pattern in AC amplitude, while the overall pattern in C-phrases is rise-rise-rise in AC amplitude. In both, C- and T-phrases, the first accent command is low in amplitude, followed by a significantly higher medial AC in both instances – in the case of C-phrases the increase is 24% (AC medial position compared to AC first position), while in the T-condition, the increase is only 15%. The pattern that follows after the medial AC is different for both prosodic phrase types. If the same calculations are made for each dialect, we find a rise-fall-rise pattern in T-phrases for all dialects except for the VS variety, where a rise-rise-fall pattern is found. In C-phrases, all dialects demonstrate rise-rise-rise patterns, not including the BE group, here a rise-rise-rise pattern is prevalent.

In both, C- and T-phrases, the performed chi-square test of independence showed significant differences between first, medial, penultimate, and ultimate AC positions in the relative proportions of falling and rising ACs (T-phrases: \( \chi^2(3, 3281) = 164.4, p < .0001 \), C-phrases: \( \chi^2(3, 6510) = 226, p < .0001 \)). In the T-condition, 28% of the first ACs are falling ACs, while in the ultimate position, only 9.4% are falling ACs. The same yet less distinct, can be found in C-phrases. All dialects show relatively similar distributions.
3.2.4. Timing

In this study, only the timing of rising ACs is investigated, since they make up 87% of all the AC movements. This timing is expressed by the distance between T1 and segment onset ($T1_{dist}$). Compared to T-phrase, C-phrase demonstrate significantly later rises (t-test $t(9578) = -4, p < .0001$). Figure 8 pictures the ANOVAs on the different AC positions’ $T1_{dist}$ (msec), according to prosodic phrase type. Both ANOVAs showed significant effects (T-phrase: $F(3, 2763) = 4, p < .007$, C-phrase: $F(3, 5782) = 10, p < .0001$).

![Figure 8](image)

Figure 8: ANOVAs on the AC positions’ $T1_{dist}$ (msec), according prosodic phrase type.

Figure 8 indicates that in T-phrase, penultimate ACs show the earliest rises, before segment onset, while ultimate and particularly first ACs rise after segment onset. If the same ACs show rises before segment onset, while ultimate and after segment onset. In C-phrases, medial and penultimate ACs show rises before segment onset, while ultimate and particularly first ACs rise after segment onset. If the same calculations are performed for each dialect, we find that in T phrases, the dialects primarily differ in the $T1_{dist}$ values of the first AC position (ZH showing the earliest rises, VS the latest), while in phrase-final positions the dialects behave similarly. In C-phrases, most dialectal variation occurs in first and penultimate AC position (GR rising the earliest, ZH the latest).

4. Discussion

The analyses conducted on Swiss German spontaneous speech confirmed significant differences between C- and T-phrases for every model parameter. The high magnitude in C-phrases as well as their short durations may be a typical feature of spontaneous answers. The informant starts at a higher F0 in anticipation of the following series of IPs and the gradual declination of F0 throughout the entire answer [cf. 12]. Due to speech disfluencies, frequently modeled with single PCs, as well as online speech planning during the articulation process, comparatively shorter prosodic phrases may be triggered.

The high AC amplitudes as well as the long AC durations in C-phrases, too, reflect high-level F0 movements. However, this finding need not necessarily suggest that phrase-medial or phrase-final segments are more locally F0-prominent. Rather, analyses of the modeling and the data in the present corpus imply that AC length and AC amplitude are more distinct in these positions so as to compensate for the steeply falling slope of high-magnitude phrase commands in C-phrases. The high frequency of falling ACs in phrase-initial position was, at a first glance, somewhat surprising, since falling ACs were particularly anticipated in phrase-final positions in T-phrases. Again, an informal check of the data reveals that the informants habitually responded to the follow-up questions of the interviewer with affirmative “yes” or “no” answers, which predominantly exhibit local F0 falls in phrase-initial position. The comparatively late $T1_{dist}$ values in the C-phrases possibly reflect the occasional list-like character of the elicited speech. In answers to questions such as “What do you plan to do after the Baccalaureate?”, the subjects frequently resorted to naming lists, thus giving rise to such F0 patterns.

The above findings also show that the dialects differ as to how C- and T-phrases are marked globally as well as locally. Each dialect exhibits different patterns in at least one of the model parameters. For instance, in T-phrases, the GR are notable with the most distinct declination and the shortest prosodic phrases. The BE group indicates a distinct pattern of AC duration in T-phrases, with a decrease-increase-increase pattern from one AC position to the next. The VS exhibit a different pattern in AC amplitude from the other groups in T-phrases. As for the timing parameter $T1_{dist}$, the ZH group illustrates the earliest rises in first ACs of T-phrases.

5. Conclusions

In this study, significant overall differences between the F0 contours of C- and T-phrases of Swiss German have been analyzed and described. In every model parameter, the two prosodic phrase types indicated significant differences. In the same vein, phonetic differences between the dialects with regard to marking C- and T-phrases intonationally were highlighted. In order to establish the perceptual weight of these findings, the confirmed differences in production shall be tested by perception experiments in future work. A highly beneficial feature of the command-response model is the easy retrieval of a synthesis of the generated F0 contour, which will facilitate the testing procedure.

6. Acknowledgements

AL is grateful to the Swiss National Science Foundation.

7. References