REAL-TIME MANIPULATION OF THE F_0-CONTOUR IN SYNTHETIC SPEECH USING THE FUJISAKI MODEL

Simon Stone, Konrad Schulze, Peter Steiner, Peter Birkholz

Institute of Acoustics and Speech Communication, Technische Universität Dresden
simon.stone@tu-dresden.de

Abstract: In this paper, we propose a system that allows the user of a real-time speech synthesizer to directly manipulate the F_0 contour of an utterance on-line and in real-time. The intonation is generated by the Fujisaki Model, which creates the F_0 contour based on accent and phrase commands that the user needs to trigger. These input commands to the model can be generated by the user with the buttons of a wireless Mycestro 3D mouse. To evaluate the usability, a study with 16 subjects was conducted and 10 monotone sentences were manipulated in real-time using the proposed system. The results show that the majority of users were able to produce a natural sounding intonation.

1 Introduction

In human speech, numerous prosodic features encode diverse information ranging from the speaker's intention (e.g., [1, 2]) to their emotional state (e.g., [3, 4]). The absence of prosody in a synthesized utterance therefore immediately degrades the perceived naturalness and thus the quality of the synthesis. One major prosodic feature is the change of the fundamental frequency F_0 over time. Current text-to-speech systems (e.g., MARY TTS [5]) derive the intonation from a combination of prosody rules based on parts-of-speech tagging and punctuation information. While this technique yields satisfying results, it requires the use of a labeling system (e.g., the German Tones and Break Indices (GToBi) [6]) to mark up the text or phonetic representation of the desired utterance. However, in systems where the synthesis is directly driven by acoustic or articulatory features, this information is not available unless explicitly passed as an additional feature. In the ongoing research efforts at our lab to develop a closed-loop articulatory synthesizer driven by articulatory measurement data from the anterior mouth cavity that can serve as a voice prosthesis, no laryngeal information and, by extension, no F_0 data is available. Because our system does not use any text or linguistic representation of the articulated speech, rule-based or parts-of-speech approaches are immediately disqualified. Instead, the user of the system has to provide the time-varying F_0 contour manually, on-line and in real-time. The direct approach to obtain the contour from the user would be to let them "direct" the speech like a director would lead a choir or orchestra, using finger and hand gestures corresponding to pitch accents or tone height. In fact, systems based on this approach exist in the context of the so-called "performative voice synthesis" (e.g., [7]). As the name of these systems suggest, they are however exclusively used in artistic performances and for educational or edutainment purposes. Because of the constant cognitive load of directly controlling the F_0 contour, we instead propose to employ a generative intonation model that only requires the user's attention at critical moments during an utterance or phrase.

2 F_0 models for intonation control

Numerous F_0 models exist and among the most commonly used in speech synthesis are the Tilt Model by Taylor [8], the Target Approximation Model (TAM) by Xu [9, 10], and the Fujisaki Model [11]. The Tilt Model is purely mathematical, in the sense that there are no underlying motivations from physiological processes during speech production. It is essentially a concatenation of parameterized curve segments to obtain any desired intonation trajectory. The user sets the duration, amplitude and tilt of each segment sequentially. While this allows total freedom in the creation of the intonation, there is also no way to limit this technique to guarantee realistic or at least physiologically possible trajectories. In contrast, the TAM is physiologically motivated. In [9], the authors describe their model as the simulation of "the effects of the aggregated force of the laryngeal controls". It uses syllable-based pitch targets, set by the user, that can either be constant (a static F_0 level) or dynamic (a linearly rising or falling target F_0). The pitch targets of an utterance are concatenated and then passed to the model, which generates the intonation curve by asymptotically approximating the target using an exponential function. By imposing reasonable limitations onto the parameters, the model will therefore always generate realistic intonation curves and artifacts can mostly be avoided. However, both the Tilt and the Target Approximation Model are based on sequentially concatenating segments (of the F_0 curve directly or of the F_0 targets, respectively). In a real-time system, this demands a lot of planning and precise control by the user, since they have limited possibilities to correct themselves once the segment has been parameterized.

The Fujisaki model, on the other hand, is superpositional in nature. The motivation for this parallel approach, according to Fujisaki [12], is that the variation of the F_0 in speech is caused by the cricothyroid muscle moving the thyroid cartilage, which in turn changes the tension of the vocal folds that are attached to it and, consequently, the F_0. This movement has two degrees of freedom (translation and rotation) and thus can be described by two components that are independent of one another: a phrase component and an accent component. These components are the responses of critically-damped second-order lowpass filters to an impulse (the phrase command) as given by

$$G_p(t) = \begin{cases} \alpha^2 t \mathrm{e}^{-\alpha t} & t \geq 0 \\ 0 & t < 0, \end{cases} \tag{1}$$

where α is the time constant of the phrase component, or to a stepwise function (the accent command) as given by

$$G_a(t) = \begin{cases} \min(1-(1+\beta t)\mathrm{e}^{-\beta t}, \gamma) & t \geq 0 \\ 0 & t < 0, \end{cases} \tag{2}$$

where β is the time constant and γ is the ceiling level of the accent component. The generated components are summed up and then added to a logarithmic base frequency to calculate the final, logarithmic F_0. As the example in Figure 1 shows, even complex F_0 contours can be generated with just a few commands.

Table 1 summarizes the properties of the three models introduced above. Even for a simple contour, the curve segments in the Tilt Model are too complicated and unintuitive for the user to parameterize in real-time and the model's non-physiological background may lead to very unnatural sounding contours, when non-optimal parameters are chosen. The Target Approximation Model should generally produce natural sounding contours and may be a good choice for simple contours (e.g., a continuously declining F_0), but becomes much more difficult to handle in real-time if used to generate more complex contours involving accents. The Fujisaki

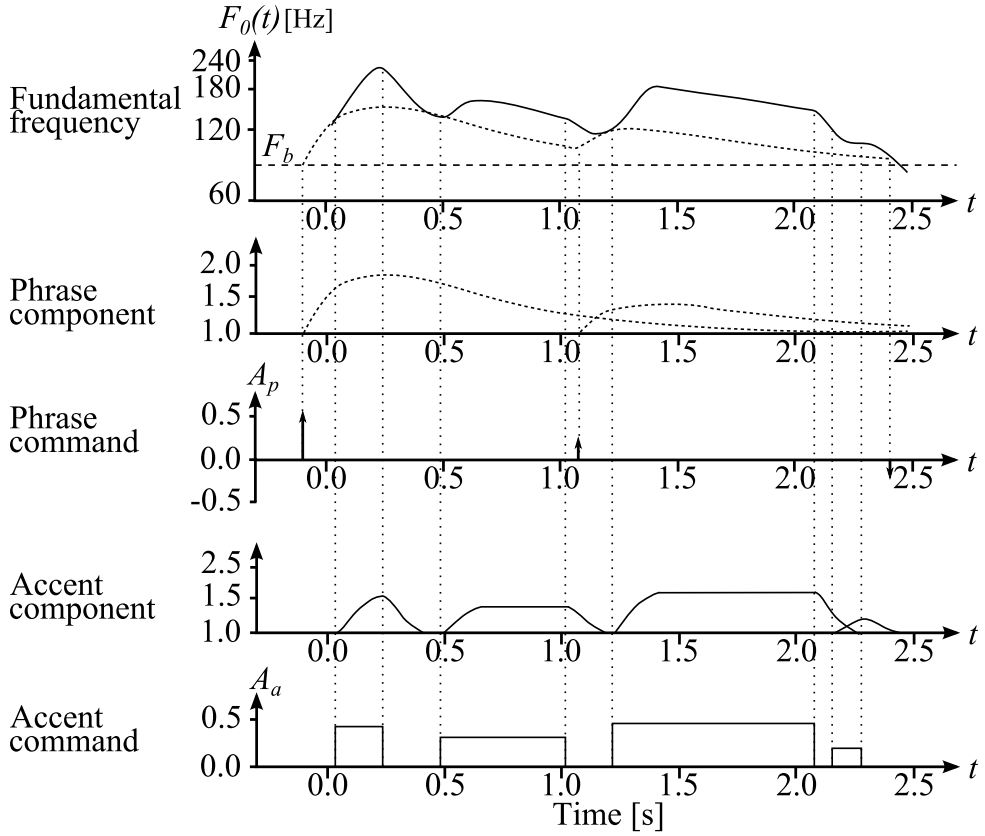

Figure 1 – The Fujisaki intonation model (figure recreated from [12]): The final F_0 contour is a superposition of a phrase component and an accent component. The phrase component is generated as the response of a critically-damped second-order lowpass filter to sequence of (weighted) impulses (the phrase commands A_p) and the accent component is the response of another critically-damped second-order lowpass to a stepwise function of varying height and width (the accent commands A_p). The components are summed up and added to a base frequency F_b in the logarithmic domain to obtain the final contour.

model, however, only needs a single parameter for a basic, declining contour and only two more for each accent. More detailed contours can easily be generated by superimposing simple contours. Therefore, it is the most suitable for the purpose of a real-time intonation generator with minimal cognitive overhead for the user.

	Tilt Model	**Target Approximation Model**	**Fujisaki Model**
Motivation	purely mathematical	physiological	physiological
Elements	pitch events	syllables	phrases and accents
Contour generation	sequential	sequential	superpositional
Parameters	3 (per element)	3 (per element)	2 per phrase, 3 per accent

Table 1 – Comparison of three intonation models commonly used in speech synthesis.

3 The Wearable Intonation Generator

To evaluate the feasibility of manipulating the intonation of an utterance in real-time, we developed a software called "Wearable Intonation Generator" (WIG) using the C++ library wxWidgets (version 2.8.12, `www.wxwidgets.org`). The software consists of a graphical user interface (see Figure 2) that allows the user to load a wave file into the program buffer. After pressing the

"Play Sound" button, the file is played back and the user can manipulate the intonation contour using the Fujisaki model, which is then impressed on the played back utterance in real-time using a Time Domain Pitch-Synchronous Overlap and Add (TD-PSOLA) algorithm. The control of the Fujisaki model is further simplified by setting the parameters α, β and γ and the phrase and accent command magnitudes for the entire utterance using the settings tab. Because the Fujisaki model is superpositional, the user can reclaim some of the freedom that is lost due to the static parameters by "stacking" several commands, which is not possible with the other models. While this restricts the shape of the components that can be generated, it reduces the entire control of the intonation generation to setting the timing of the commands using only two buttons: one to trigger a phrase command (an impulse) and one to trigger an accent command (push to step up, hold, release to step down). The software supports two input modalities to use as these controls: the keyboard (Spacebar for phrase, Ctrl for accent commands) or the mouse (left button for phrase, right button for accent commands). Since the system is intended to be used as a component in a wearable speech synthesis system, WIG also supports the use of the wireless Mycestro 3D mouse (`www.mycestro.com`), which is a small device that is strapped to one of the user's index fingers and communicates with the PC via Bluetooth. The Mycestro mouse supports three buttons and a scroll wheel, of which only two buttons (left and right) are needed to control the WIG.

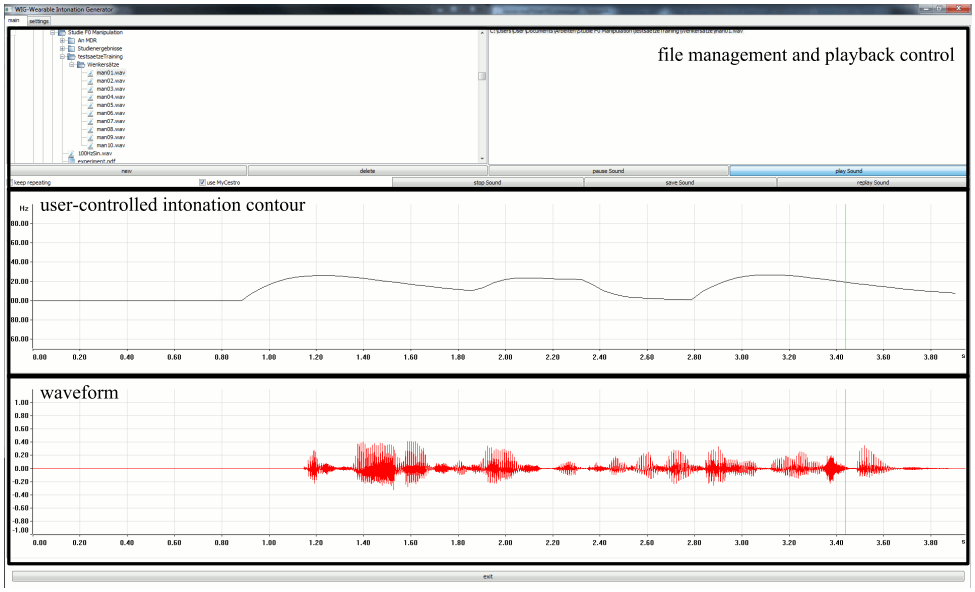

Figure 2 – Graphical user interface of the Wearable Intonation Generator. The settings tab hides the parameter settings for the Fujisaki model. During playback of the wave file, the user can generate phrase and accent components by setting the timing of the corresponding commands using the keyboard or a wearable 3D mouse. The F_0 of the wave file is manipulated on-line using a TD-PSOLA algorithm.

4 Usability study design

Even though only two buttons are needed to create even complex F_0 contours, the task to give the commands on-line and in real time is unusual for the user and requires a usability test. We therefore designed a study with 16 subjects (native-level German speakers, age 21-30) who were asked to manipulate the F_0 contours of 10 German sentences (based on the Wenker sentences used in [13], see Table 2). The sentences were recorded with a professional male speaker and their intonation flattened to a static F_0 of 100 Hz using the software Praat [14]. This approach was chosen over a (unit selection or articulatory) synthesis of the sentences with a constant F_0

because the introduction of additional unnaturalness due to synthesis artifacts was to be avoided. It also provided a truly natural sample for reference during the rating part.

	German	English translation
1.	Das Feuer war zu heiß, die Kuchen sind ja unten ganz schwarz.	The fire was too hot, the cakes are all black on the bottom.
2.	Wem hat er denn die neue Geschichte erzählt?	Whom did he tell the new story?
3.	Ihr dürft nicht solche Kindereien treiben!	You should not horse around!
4.	Das war recht von Ihnen!	You did good!
5.	Ich bin mit den Leuten da hinten über die Wiese ins Korn gefahren.	I drove with the people over there across the meadow into the field.
6.	Wir sind müde und haben Durst.	We are tired and thirsty.
7.	Es hört gleich auf zu schneien, dann wird das Wetter wieder besser	It is going to stop snowing soon, then the weather will improve again.
8.	Wie viel Pfund Wurst und wie viel Brot wollt ihr haben?	How many pounds of sausage and how much bread do you want?
9.	Ich verstehe euch nicht, ihr müsst ein bisschen lauter sprechen.	I don't understand you, you have to speak a little louder.
10.	Er ist vor vier oder sechs Wochen gestorben.	He died four or six weeks ago.

Table 2 – List of German test sentences used in the usability study and their English translation for reference.

In preparation of the experiment, each subject was asked to familiarize themselves with the controls of the software using it with a standard desktop mouse on their own computer and by manipulating a continuous 100 Hz sine tone. During the experiment, each subject used the lab computer and the Mycestro 3D mouse. In order to avoid mistakes due to the use of this unusual input device, each subject did a little exercise where they had to repeatedly click with the 3D mouse into a specific cell of a spreadsheet. Once they were comfortable with the handling of the device, each subject was presented with one of the 10 monotone sentences and was asked to manipulate the intonation in real-time during playback to make it sound more natural. If they were not satisfied with the result of a manipulation, they were allowed to try again with the same sentence. Once the subject was content with the result or after a maximum allowed manipulation time of three minutes, the last generated contour and the corresponding manipulated audio file were saved and the experiment continued with the next sentence. The order in which the sentences were presented was randomized for each subject. After all 160 manipulations (16 subjects times 10 sentences) had been made, the entire set of 160 manipulated audio files plus the 10 original recordings (with a natural intonation) and the 10 recordings with a flattened intonation was rated by each subject on a naturalness scale from 1 (totally unnatural) to 5 (totally natural).

5 Results

For each subject, 10 manipulations were rated. Calculating the global mean across all 10 manipulations would potentially yield large standard deviations (SD), since some of the sentences may have been easier to manipulate than others. So instead of the global mean and SD, the mean and SD of the ratings of each manipulation was calculated. Since every manipulation produced by a subject is a manifestation of that subject's ability to use the system efficiently, we regard this proficiency as a process and the manipulation as a realization from that process. To characterize each subject's ability to produce natural sounding contours we then calculated the naturalness score ν by combining the means and standard deviations across the manipulated samples from each subject by iteratively multiplying the corresponding (assumingly) Gaussian density functions. The result of each multiplication of a distribution with a mean μ_i and a SD σ_i and a distribution with a mean μ_j and a SD σ_j is again a Gaussian density function with mean

μ_{ij} and SD σ_{ij} according to:

$$\mu_{ij} = \frac{\mu_i \sigma_j^2 + \mu_j \sigma_i^2}{\sigma_i^2 + \sigma_j^2} \qquad \text{and} \qquad \sigma_{ij} = \sqrt{\frac{\sigma_i^2 \sigma_j^2}{\sigma_i^2 + \sigma_j^2}} \tag{3}$$

The results of the study are summarized in Figure 3. Compared to the sample with a flat intonation, only subject 14 was generally not able to improve the naturalness of the intonation with our system. While a distinct gap remains in the perceived naturalness between the original recording and even the manipulations made by the best subject, the majority of the subjects (10 out of 16) achieved a naturalness score of more than three, which is considered natural. As illustrated by Table 3, there is only a very weak positive correlation between the average time to settle on a contour and its naturalness, and a somewhat stronger, but still weak, negative correlation between the average number of attempts and the naturalness.

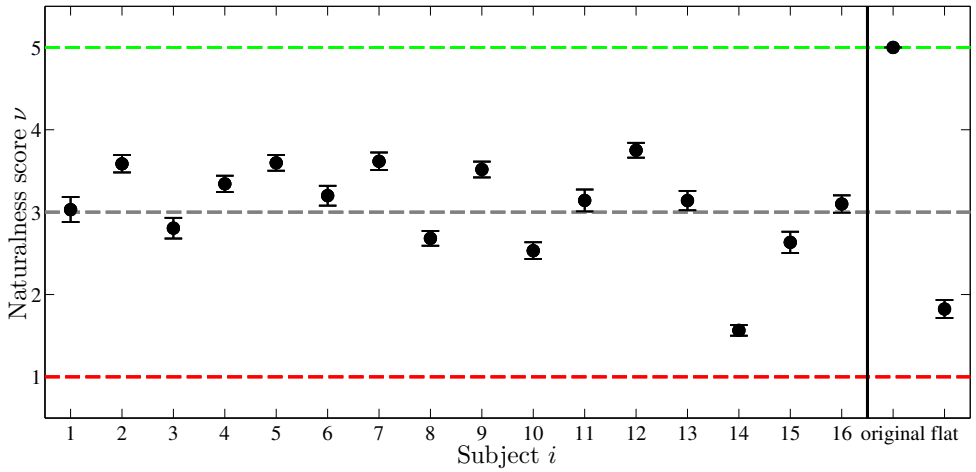

Figure 3 – Mean (dots) and variance (whiskers) of the perceived naturalness ν (rated by all subjects) of the 10 samples generated by the respective subject i. The original sample was a natural recording, the flat sample was the same recording manipulated to a flat 100 Hz intonation.

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Mean manipulation time [min:s]	n/a	n/a	1:14	1:28	0:39	1:06	1:28	0:47	1:00	1:00	1:11	1:19	0:55	1:16	1:11	0:57
Mean number of attempts	n/a	n/a	11	9.1	5.1	9.8	12.4	7.7	10.7	10.7	16.2	19.9	13.9	21.6	17.3	5.8
Naturalness score	3.03	3.59	3	3.34	3.6	3.2	3.6	2.7	3.5	2.5	3.1	3.75	3.14	1.56	2.6	3.1

Table 3 – Average amount of time and mean number of attempts each subject needed to settle on a contour. The time and number of attempts was not tracked for the first two subjects. The correlation coefficient between the average manipulation time and the naturalness score is $\rho_t = 0.012$ and the correlation coefficient between the number of attempts and the naturalness score is $\rho_n = -0.38$.

6 Summary and conclusion

We introduced a system consisting of a wearable input device (Mycestro 3D mouse) and a PC software that allows the real-time on-line manipulation of the intonation of pre-recorded sentences with flat intonation. The system's usability was evaluated with a small-scale user study and the results show that the majority of the users were generally able to produce natural sounding F_0 contours. However, the users were given an indefinite number of retries and a rather

long time period to manipulate the short sentences. Future studies should therefore examine, how the naturalness is affected by stricter limitations, since the final application is supposed to be in a wearable speech synthesis system, where usually only a single attempt for each sentences is possible. Another future study should also examine, how subjects improve over time as they train on one set of sentences with an indefinite number of attempts and are then tested on a different second set with only one attempt per sentence. It is also of interest to see if the users' proficiency can be improved by teaching them the theory behind the Fujisaki model. In the presented study, the users were not told anything about the underlying concept of distinct phrase and accent components to allow them more freedom in how they use the system. As shown by Figure 4, a basic understanding of this concept may lead to better results.

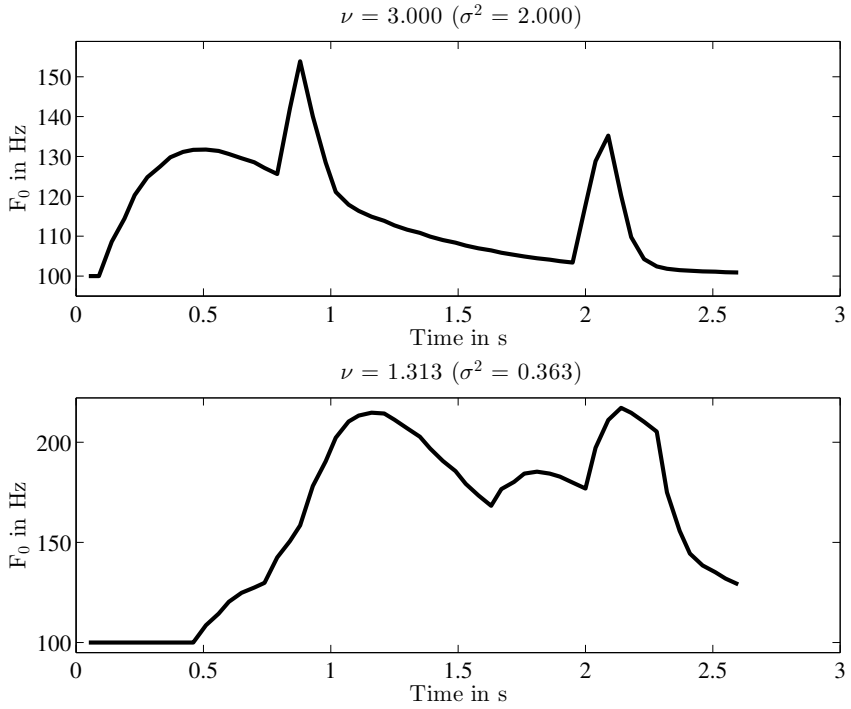

Figure 4 – Example F_0 contours generated by the highest-scoring subject (above) versus the lowest-scoring subject (below). Apparently, both subjects had a similar target contour in mind (one phrase with two accents) but subject 14 seemingly had trouble with both timing and choice of the commands. Future work should investigate if these problems can be mitigated by further and supervised training.

Another possible improvement could be to use a declining base F_0 or even a phrase component as a base. This would further reduce the user's workload to triggering the accents only, which may be easier to accomplish because of the more immediate response to a command (as opposed to the comparatively slowly rising response to a phrase command). This is especially advantageous for the target audience of a speech prosthesis, among which cognitive impairments are likely to occur. Lastly, because the present study did only consider the general naturalness (a fairly abstract measure), another experiment should examine to what extent the subjects are able to intentionally convey information using the artificial intonation (e.g., stressing specific words or syllables to resolve ambiguities).

7 Acknowledgments

The authors would like to thank Susanne Drechsel of the Martin Luther University Halle-Wittenberg for the recordings of the test sentences. This project was funded by the German Federal Ministry of Education and Research (BMBF), reference number 13GW0101B.

References

[1] SNEDEKER, J. and J. TRUESWELL: *Using prosody to avoid ambiguity: Effects of speaker awareness and referential context*. *Journal of Memory and Language*, 48(1), pp. 103–130, 2003.

[2] CUTLER, A., D. DAHAN, and W. VAN DONSELAAR: *Prosody in the comprehension of spoken language: A literature review*. *Language and Speech*, 40(2), pp. 141–201, 1997.

[3] SCHERER, K. R., R. BANSE, H. G. WALLBOTT, and T. GOLDBECK: *Vocal cues in emotion encoding and decoding*. *Motivation and emotion*, 15(2), pp. 123–148, 1991.

[4] PELL, M. D.: *Influence of emotion and focus location on prosody in matched statements and questions*. *The Journal of the Acoustical Society of America*, 109(4), pp. 1668–1680, 2001.

[5] SCHRÖDER, M. and J. TROUVAIN: *The German text-to-speech synthesis system MARY: A tool for research, development and teaching*. *International Journal of Speech Technology*, 6(4), pp. 365–377, 2003.

[6] GRICE, M., S. BAUMANN, and R. BENZMÜLLER: *German Intonation in Autosegmental-Metrical Phonology*. *Prosodic typology: The phonology of intonation and phrasing*, 1, p. 55, 2006.

[7] FEUGÈRE, L., S. LE BEUX, and C. D'ALESSANDRO: *Chorus digitalis: polyphonic gestural singing*. In *1st International Workshop on Performative Speech and Singing Synthesis (P3S 2011), Vancouver (Canada)*, vol. 14. 2011.

[8] TAYLOR, P.: *The Tilt intonation model*. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*, p. 0827. International Speech Communication Association, 1998.

[9] XU, C. X., Y. XU, and L.-S. LUO: *A pitch target approximation model for F0 contours in Mandarin*. In *Proceedings of the 14th International congress of Phonetic Sciences*, pp. 2359–2362. 1999.

[10] XU, Y. and Q. E. WANG: *Pitch targets and their realization: Evidence from Mandarin Chinese*. *Speech communication*, 33(4), pp. 319–337, 2001.

[11] FUJISAKI, H.: *A model for synthesis of pitch contours of connected speech*. *Annual Report, Engineering Research Institute, University of Tokyo*, 28, pp. 53–60, 1969.

[12] FUJISAKI, H.: *Information, Prosody, and Modeling - with Emphasis on Tonal Features of Speech*. In *Speech Prosody 2004*, pp. 1–10. International Speech Communication Association, 2004.

[13] BOCK, D., B. GANSWINDT, H. GIRNTH, S. KASPER, R. KEHREIN, A. LAMELI, S. MESSNER, C. PURSCHKE, and A. WOLANSKA: *Regionalsprache.de (REDE). Forschungsplattform zu den modernen Regionalsprachen des Deutschen*. Forschungszentrum Deutscher Sprachatlas, Marburg, 2008 ff.

[14] BOERSMA, P.: *Praat, a system for doing phonetics by computer*. *Glot International*, 5(9/10), pp. 341–345, 2001.