Modeling Prosody in a Cross-Language Perspective

Hansjörg Mixdorff

Berlin University of Applied Sciences
Luxemburger Str. 10, 13353 Berlin, Germany
mixdorff@tfh-berlin.de

ABSTRACT

Although it has been shown for many languages that the Fujisaki model is in principle capable of decomposing *F0* contours in these languages, problems remain as to the automatic extraction as well as the interpretation of parameters extracted with respect to their phonological relevance. The claim is made in the current article that the Fujisaki model parametrization used must be rooted in the linguistic system of the particular language, and that this consideration should always be given priority over simple fitting accuracy. To this effect, this article presents settings and results from a number of more recent studies by the author and his co-workers in various languages, such as German, Finnish, Thai and Vietnamese.

1. Introduction

It is widely agreed that the Fujisaki model [1] presents one of the most interesting approaches to modeling *F0* contours. As has been shown in numerous studies on a large variety of languages, *F0* contours can be faithfully decomposed into two components: The slow changing *phrase component* and the fast changing *accent component*. We should remind ourselves that these original terms refer to phrases and accents in the common Japanese, a language with an accent system that is closely linked to the moraic structure of the language. The accent type of a lexical word is characterized by the index of the mora after which a distinct fall in the *F0* contour occurs. There exist, however, also so-called unaccented words which exhibit a very flat *F0* contour and no distinct fall. These connect with accented words forming a longer prosodic group. A phrase component is usually associated with such a prosodic group which typically contains clitically linked function words on its right edge. Phrase and accent components are the output of two critically-damped linear filters and the responses to so-called input commands, impulse-wise accent phrase commands and step-wise accent commands. In later works, Fujisaki and his co-worker attempted to develop a physiological interpretation of the model commands being associated with neuro-motor control of an intrinsic muscle of the larynx, the crico-thyroid muscle which acts as an antagonistic muscle to the vocalis muscle alongside the glottis[2]. It is the latter interpretation which is also unique to the Fujisaki model.

Having said so much, one could think that apparently the model would be the right choice for *F0* analysis as well as synthesis. Still, working TTS systems applying the model can be counted on one hand, and general studies on prosody applying the model are almost non-existing if one counts out the works by the current author. So what is it that seems to scare off researchers from using the model?

In the first place, since the model has a physiological motivation, it does not come with a set formula of how to relate its parameters to linguistic and para-linguistic units and structures. As a matter of fact, accent and phrase component have their linguistic correspondences, namely Japanese tone accents and bun-setsu (prosodic words). These, however, do not necessarily coincide with similar structures in European languages, not to mention tone languages. This means that a linguistically feasible interpretation must be developed for each language individually. It should be noted, however, that this is also true for the widely used ToBI system since the required label sets are strictly language-specific[3].

Secondly, parameters, that is, phrase and accent commands must be derived from the natural *F0* contour in an error-prone AbS procedure. As has been shown by Fujisaki and his co-workers as well as the current author, reasonably stable algorithms for doing this decomposition can be developed, but they require a high degree of expert knowledge and post-processing [4][5]. Their present degree of robustness can probably not compete with similar algorithms as the one for Paul Taylor’s Tilt model [6] or the Codebook Vector representation developed by Gregor Moehler [7], for instance, which basically break down the *F0* contour into a sequence of shapes. The great advantage of the Fujisaki model providing a decomposition of the *F0* contour into sequences of events governing different domains, that is, prosodic phrases and accents which can then be related to the corresponding linguistic units, proves to be the largest difficulty with respect to parameter estimation, namely the separation of phrase and accent components. Furthermore, an infinite number of commands may provide an infinitely accurate fit, but is hard to interprete meaningfully. It is therefore the trade-off between fitting accuracy and meaningful parametrization that requires a large amount of experience. As a consequence, there are definitely more straight forward approaches to parameter extraction for speech recognition and synthesis on whose output statistical models can be easily trained. If one, however, is more interested in understanding how prosody works - in production as well as perception - the Fujisaki model should still be regarded as an attractive choice.

Needless to say, a suitable way of parametrizing *F0* contours of a particular language must be based on the linguistic properties of that specific language. In brief, a mapping between phrase and accent commands and linguistic units and structures needs to be established. This mapping should account for phonological differences of the specific languages. It also needs to be decided whether a specific languanges requires negative accent commands.

In the case of languages such as English or German, the notion of 'accent' is a well established property. Accents are most likely to occur on syllables which carry the lexical stress of that word. Depending on certain accentuation rules of German and in contrast to Japanese, accents can be subject to deletion, especially in noun-verb combinations where depending on the semantic context either the noun or verb can be deaccented. Hence, there is a close relationship between an accent command yielded in the analysis with the Fujisaki model and an accented syllable.

In order to link Fujisaki model commands to the rhythmical structure of an utterance, these are connected to syllable timing. In the framework of IGM, the integrated prosodic model developed by the author [8], accent command timing is related to the corresponding accented syllable bounds. Furthermore, phrase commands are related to the onset of the first syllable of the phrase which they precede. Depending on the language, the relationship for *T1* can be established with respect to the syllable onset or the onset of the syllable rhyme.

The following four sections briefly present the specific approaches developed to modeling *F0* contours for German, Finnish, Thai and Vietnamese. The mapping of Fujisaki model commands on linguistic units and structures is discussed, especially the function of accents commands (tone commands in the case of Thai and Vietnamese) in these languages. Each section presents the short discussion of a production/perception experiment exploring a particular phonological distincion in the language concerned. These experiments follow a similar strategy: A corpus of segmentally identical, but prosodically different utterances is created. This corpus is then analysed using a Fujisaki model formulation developed for the specific language. Parameters established as pertaining to a certain phonological function are normalized and averaged and subsequently used for determining so-called center-stimuli for the conditions which they represent. The parameters are then employed for creating resynthesis stimuli supposed to convey the intended phonological function. Except for the section on Vietnamese where the perception experiment only concerned simple tone identification, intermediate stimuli between the center-stimuli were created in order to determine which prosodic parameters make subjects' decisions shift from one category to another.

2. Modeling Sentence Mode in German

According to the works by Isačenko & Schädlich [9] and Stock & Zacharias [10], a given *F0* contour can be described as a sequence of linguistically motivated tone switches, major transitions of the *F0* contour connected to accented syllables, or by so-called *boundary tones* before prosodic boundaries. Tone switches can be thought of the phonetic realization of phonologically distinct intonational elements, so-called 'intonemes'. In the original formulation by Stock, depending on their communicative function, three classes of intonemes are distinguished, namely the N↑intoneme ('non-terminal intoneme' at phrase-medial accents, rising tone switch), I↓ intoneme ('information intoneme' at declarative-final accents, falling tone switch), and the C↑ intoneme ('contact intoneme' associated with question-final accents, rising tone switch). Hence intonemes in the original sense mainly distinguish sentence modality, although there exists a variant of the I↓ intoneme, I(E)↓ which denotes emphatic accentuation and occurs in contrastive environments, for instance. Intonemes – except for I(E)↓ which is governed by the context of an utterance - are predictable by applying a set of phonological rules to a string of text as to word accentability and accent group forming.

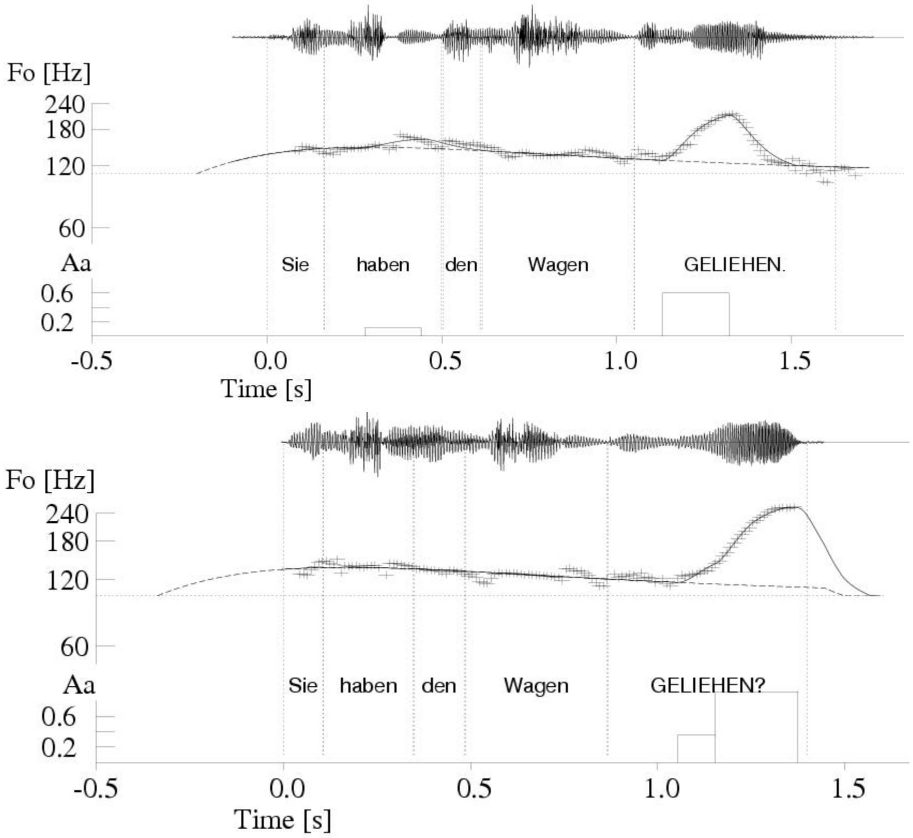


Figure 1: Two utterances with different sentence mode (statement, top, question, bottom) of the sentence ‘sie haben den Wagen geliehen’.

In order to quantify the interval and timing of the tone switches with respect to the syllabic grid, accented syllables and syllablese bearing a boundary tone are related to accent commands. In a production/perception experiment [11], 18 native speakers of German uttered the sentence ‘sie haben den Wagen geliehen’- *They have rented the car* - in varying contexts, namely as a single-phrase question or statement, respectively, or the first phrase in a two-phrase utterance. Narrow focus was placed on the word ‘geliehen’. *F0* contours of the utterances where parametrized as shown in Figure 1. By relating the accent commands associated with the last word prototypical configurations were established for the conditions question (accent command pair with relative high amplitudes), statement (single short accent command early in the word) and non-terminal (single accent command late and extending to the phrase boundary). These normalized configurations were used to create resynthesis stimuli representing the three different conditions. Whereas statement vs. non-terminal differed as to their *T1/T2* alignment, non-terminal and question differed as to accent command splitting. Starting from the three center stimuli for the categories intermediate stimuli were created and presented to 22 native speakers of German. The resulting judgments are displayed in Figure 2. The center stimuli (marked with black dots) are unanimously identified with their respective categories. The solid lines represent lines of equi-probability for subjects’ judgments, and the white dots the locations of the intermediate stimuli. As can be seen in the left graph, non-terminal intonation is mostly identied by the later offset time *T2* as compared with statement intonation.

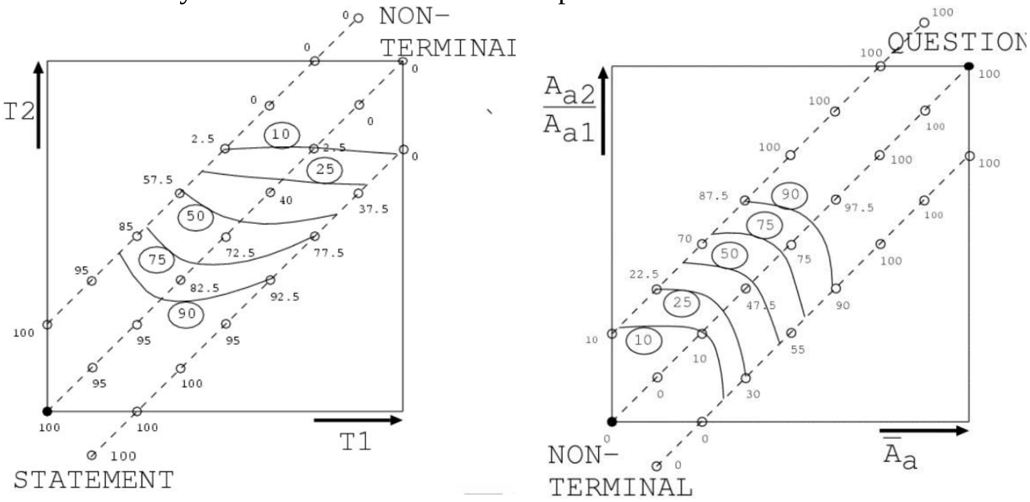


Figure 2. Results of perception experiment on sentence mode (left: statement vs. non-terminal, right: non-terminal vs. question).

The right graph shows that question intonation is identified either by the higher mean accent command amplitude of the two consecutive commands or by the amount by which the accent commands differ in amplitude. This splitting of the accent command is due to the so-called ‘question-final rise’ which is actually a high boundary tone falling the rise associated with the accent on the syllable ‘lie’.

3. Modeling Focus in Finnish

Finnish has a fairly free word order, rich morphology with suffixation, and enclisis, as well as a relatively large number of grammatical cases, a set of features which is typical for an intonationally falling language. Due to suffixation and enclisis, the lexical morphemes are at the beginnings of the words leading to a state in which the lexical stress is invariably on the first syllable of the word. The basic intonation shape in Finnish is a falling shape with an accent on basically all content words. Finite verbs are usually less prominent than nouns and are sometimes altogether unaccented. Finnish questions are typically marked only by lexical means (by interrogative particles) which has lead most researchers (see Iivonen [12], for instance) to the conclusion that there is no interrogative intonation in Finnish. Continuation is typically signaled by level intonation and finality by a sharp fall into the bottom of the speakers fundamental frequency range, which usually causes a creaky or whispery voice during the last unstressed syllables of an utterance. It has been shown in [13] that lexical accent syllables are aligned with accent commands whose amplitude increases in the presence of narrow focus. Different from German, tone switches - at least in the read material analyzed in this study - are generally falling and do not undergo complete deletion in pre-focal and postfocal position.

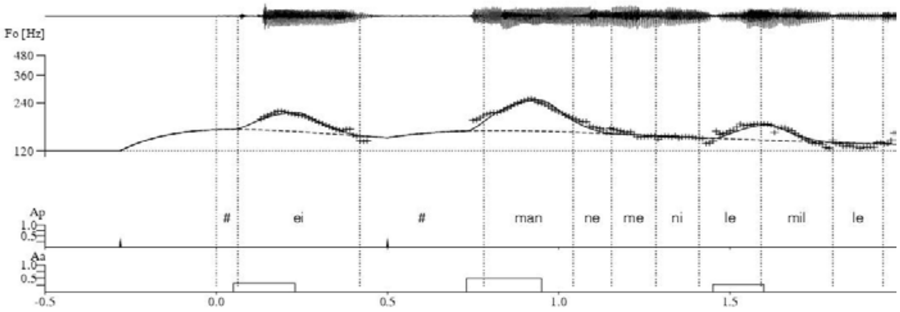


Figure 3. A typical Finnish utterance ‘Manne meni Lemille’ with double contrast on the words ‘Manne’ and ‘Lemille’ produced by a female speaker of Finnish and analyzed with the Fujisaki model.

For our production/perception study [14] we used the sentence ‘Manne meni Lemille’ (Manne went to Lemi) which permits four possible interpretations with respect to focus:

- broad
- narrow focus on ‘Manne’
- narrow focus on ‘Lemille’
- double contrastive, narrow focus on both ‘Manne’ and ‘Lemille’

A typical utterance spoken by a female speaker and analyzed with the Fujisaki model can be seen in Figure 3. The figure shows the actual waveform of the utterance, the modeled as well as the actual contours (solid line and +-marks, respectively). The figure also shows syllabic labels of the utterance and the parameters for the Fujisaki model (the phrase command is actually outside of the scope of the figure, but the accent commands are conspicuous. The utterance has a double contrast on the words 'Manne' and 'Lemille').

The 12 students who participated in the test were divided into six pairs, in which one acted as the questioner whose task was to read aloud a prompt to the other participant who then produced the intended reply in the desired focus condition. Table 1 shows the typical prompt-reply pairs used in the study. The prompt-reply pairs were presented to the participants in written form on a sheet of paper.

Table 1. Typical prompt-answer pairs used to elicit the different focus conditions.

A:	Mitä sitten tapahtui?	What happened then?
B:	Manne meni Lemille.	Manne went to Lemi.
A:	Kuka meni Lemille?	Who went to Lemi?
B:	**Manne** meni Lemille.	**Manne** went to Lemi.
A:	Minne Manne meni?	Where did Manne go?
B:	Manne meni **Lemille**.	Manne went to **Lemi**.
A:	Kuulin, että Manu	meni Lemulle.
	I heard that Manu	went to Lemu
B:	Ei -**Manne** meni	**Lemille**.
	No **- Manne** went	to **Lemi**

The target utterances were segmented on the syllabic level and the parameters for the Fujisaki model were estimated manually using an interactive program with a graphical user interface. Materials from a typical speaker were selected for further analysis to serve as the basis for the perception experiment.

The data from the utterances were analyzed in order to build statistically representative stimuli for the experiment. Typical configurations exhibited one or two accent commands associated with 'Manne' and/or 'Lemille'. Accent command amplitudes for the four focus conditions are given here for a female speaker whose utterance were later on used in a perception experiment: 1) *Aa*1/*Aa*2: 0.25/0.33; 2) 0.41/0.08; 3) 0.00/0.44; 4) 0.42/0.28. Based on these results an experiment with respect to the perception of focus using resynthesized stimuli was performed. Starting from an utterance from a broad focus condition, stimuli were created with *Aa*1 and *Aa*2 ranging between .00 and .42 in steps of .07 yielding altogether 49 stimuli. The modification was only done with respect to accent command amplitudes; accent command timing as well as the phrase component were kept constant. Figure 4 shows the locations of the averaged accent command amplitude values for the four different focus conditions in onto the *Aa1/Aa2* space as stars.

The same students that had participated in the production of the utterances took also part in the perception experiment. Subjects were instructed to decide which condition they thought they heard in a forced choice test by marking their response on paper. The list of stimuli was played twice in a random order with an inter-stimulus interval of approximately 2 seconds.

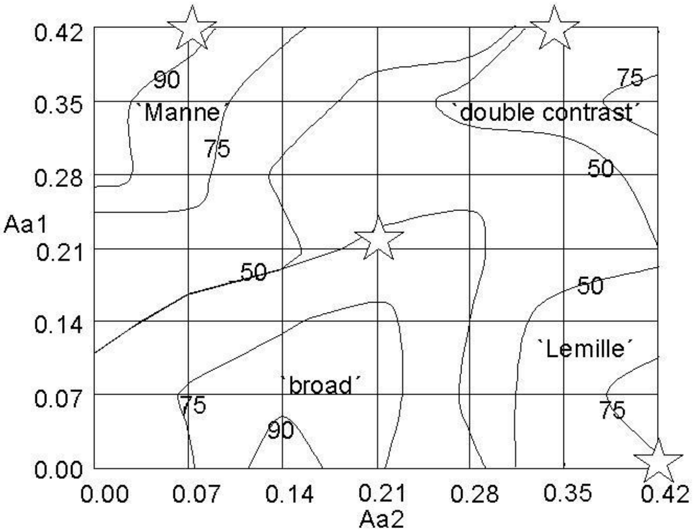


Figure 4. Equi-probability lines for the four different focus conditions depending on the accent command amplitudes of the two peaks. The stars denote the approximate values for the produced averages of the four options.

The results include, inter alia, that the second accent in the utterance must be raised by at least 2.6 semitones from the baseline to be perceived to have narrow focus on the latter word of two-accent utterance. However, the first accented word receives the perception of narrow focus with only a 1.5 semitone raise – in the case that the second word is completely unaccented. The rise of the narrowly focused item must be increased accordingly when the accent command amplitude of the other item rises. Only when both rises reach a value of approximately 3.6 semitones, are both target words are perceived as focused simultaneously. Stimuli with rises closest to the means given for the four conditions above were generally identified as belonging to the intended category, though the condition ‘broad’, apparently the default choice, covers a large triangular region in the two-dimensional accent space (see Figure 4).

With respect to statistical analysis, the accent amplitude parameter shows a clear influence on subjects' responses. Analyses of variance with the mean responses (rounded to the closest category) as grouping variable show an *F* value of 12.334 for *Aa1* and of 44.867 for *Aa2*, both at a significance level of $p<0.0001$.

The fact that in our experiment, the perception of the category ‘broad’ requires the second peak in the utterance to be lower than the first one, gives some evidence to the hypothesis that Finnish speakers and listeners normalize for the baseline declination.

4. Modeling the Syllabic Tones of Thai

The Thai language has five different lexical tones, namely three static tones, mid (0), low (1) and high (3), and two dynamic tones, falling (2) and rising (4) (tone indices commonly used given in brackets). Furthermore, a phonemic distinction exists between long and short vowels [15]. As a consequence, there exist groups of words which stand in tone/vowel quantity opposition, that is, either share the tone or vowel quantity as shown in the following example (long vowels are indicated by vowel symbol doubling):

Word	Thai script	tone	Vowel quantity	Translation
loon0	โลน	mid	long	crab louse
loon3	โล้น	high	long	to be bald
loon4	โหลน	rising	long	Great great grandson of daughter
lon0	ลน	mid	short	to singe
lon3	ล้น	high	short	to overflow
lon4	หลน	rising	short	a kind of food

Hence these groups of words present a ‘worst case’ for production as well as for perception, since in theory the correct lexical access should be possible based on the tone and vowel quantity only, if the words are uttered in isolation. In a study the segmental and tone properties of a selection of highly confusable mono-syllabic words following the example above were analyzed [16]. A speech corpus was designed which contains 17 groups of highly confusable words embedded in the carrier sentence [th@: aU] X [ma: du:], “you brought X to look at.” The carrier sentence contained mid tone syllables throughout. A semi-automatic procedure for estimating the parameters of the Fujisaki model was applied to the *F0* contours which was based on a modified version of [3], but requires a pre-segmentation of the utterance into syllables. Parameter configurations were checked and if necessary corrected. Table 2 lists averaged onset and offset times of tone commands as well as tone command amplitudes. As can be seen, the low tone as well as the rising tone require tone commands of negative polarity, whereas the mid tone can be modeled using the phrase component only.

Table 2: Average onset and offset times of tone commands with respect to the rhyme onset time in ms, and average accent command amplitudes for the five syllabic tones.

	$T11_{rel}$	$T21_{rel}$	*Aa1*	$T12_{rel}$	$T22_{rel}$	*Aa2*
mid tone	-	-	-	-	-	-
low tone	-3	220	-0.15	-	-	-
falling tone	-85	194	0.19	-	-	-
high tone	106	411	0.15	-	-	-
rising tone	-26	172	-0.19	172	417	0.10

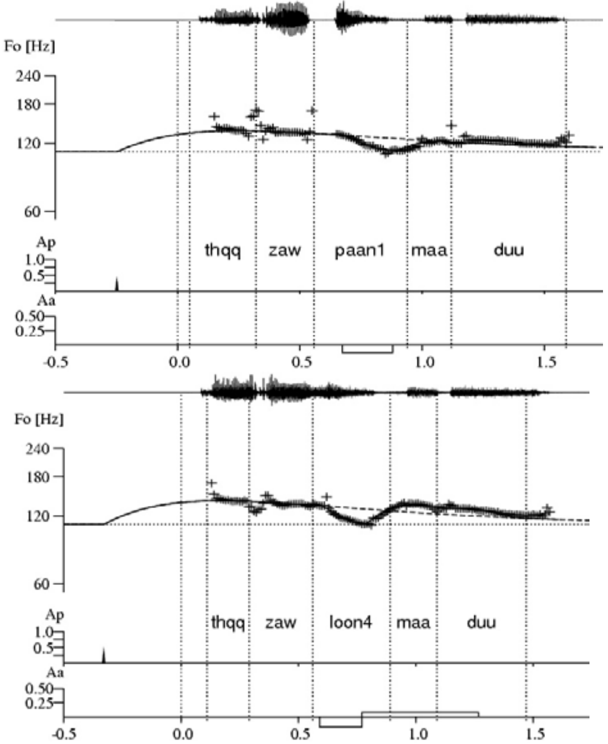
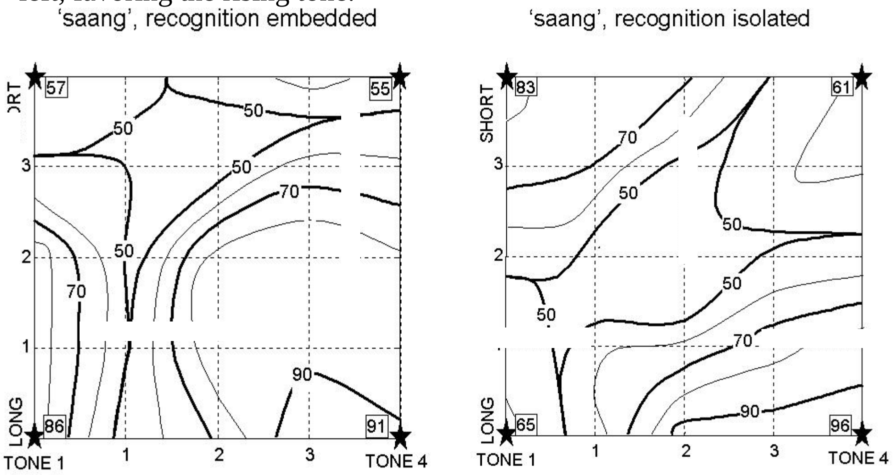


Figure 5:Examples of analysis of low (top) and rising (bottom) tone words embedded into a context of mid tone syllables.

Figure 5 displays examples of analysis for the former two tone types. Based on the result of the production data, a perception experiment was devised in which participants had to decide which member of a highly confusable word group they had perceived. In addition to the resulting center stimuli for each word, a matrix of intermediate stimuli was created. 22 phonetically untrained undergraduate and master students (14 male, 8 female) of Chulalongkorn University took part in the perception experiments.

The grid of dashed lines in Figure 6 indicates the stimuli which are located on the grid vertices. Corner stimuli in the left graph, for instance, correspond to the words 'saang1' (left bottom corner), 'saang4' (right bottom corner), 'sang1' (left top corner), and 'sang4' (right top corner). The value of 86 written in the box next to the left bottom corner means that the corner stimulus was identified as pertaining to the word 'saang1' by 86% of the subjects. The contours of 80, 70, 60 and 50% drawn around the corner stimulus indicate the decrease of the vote 'saang1' as we move away from the corner stimulus, that is, gradually move from a low tone to a rising tone, and from a long vowel to a short vowel word condition. The categorical judgment shifts where the 50 % lines around the corner stimuli coincide. There is, however, a region where neither of the four words in the highly confusable group reaches 50%. The orientation of the contours indicates how the judgment is influenced by the tone and quantity properties, that is, vertical lines suggest mainly distinction by tone, and horizontal lines distinction by vowel quantity (see 50 % line of vote 'saang1'). Words when presented in isolation are more often identified as bearing a short vowel. The regions where none of the choices reaches 50% generally increases for the isolated condition. For the long vowel condition (bottom of panels), presentation in isolation shifts the categorical 50% boundary further to the left, favoring the rising tone.


‘saang’, recognition embedded

‘saang’, recognition isolated

Figure 6: Results of perception experiment for syllable ‘sa(a)ng‘) under embedded (left) and isolated (right) conditions. The black stars denote the corner stimuli of which the respective identification rate is given in percent. Intermediate stimuli are located at equal distance on the mesh points of the grid. Lines of equi-probability are marked at 50, 60, 70, 80 and 90 %. identification rate, respectively.

This can be explained by the fact that the low tone is defined with respect to the preceding relatively higher syllable which is not present in the isolated condition. The *F0* rise which is the perceptual cue of the rising tone, however, prevails as it occurs in the long vowel. In the short vowel word condition (top of panel), this rise occurs later towards the coda of the syllable and reaches its maximum in the (not present) following syllable. Here we observe the opposite effect: The low tone is favored over the rising tone.

5. Modeling the Syllabic Tones of Vietnamese

Vietnamese is known as a monosyllabic tone language having six different lexical tones. These are (numbers indicate the indices to be used throughout this article): Level (1), sometimes also referred to as ‘mid-level’, rising (2), broken (3), falling (4), curve (5), and drop (6) tones. Tones 2-6 are marked by diacritics in the Vietnamese script which uses the Latin alphabet. The widely cited description by Thompson [17] is summarized in Table 3.As can be seen Vietnamese tones are not only characterized by distinct *F0* trajectories, but also by articulatory distinctions and the presence/absence of glottalization. Since, however, features such as voice quality and glottalization are not incorporated in the Fujisaki model, the question must be examined whether the Vietnamese tones can be reliably signaled by means of *F0* manipulation only, or whether additional control is required. In order to examine the realization of individual tones, as well as tone coarticulation, a set of 52 six-syllable utterances with varying combinations of tones was recorded by two phonetically trained native speakers of Northern Vietnamese, the Standard dialect of Vietnam, one male and one female. The utterances were composed of voiced sounds only for continuous *F0* contours, using only nasals and laterals as initial or final consonants in order to minimize microprosodic effects.

Table 3: Description of the six syllabic tones of Vietnamese.

No.	Vietnamese Name	English Name	*F0* contour	Diacritic used in writing	Additional features
1	Ngang	level	Trailing/falling	none	Laxness
2	Sác	rising	Rising	Á	Tenseness
3	Ngã	broken	Rising	Ã	Glottalization
4	Hỏi	falling	Falling	Ả	Tenseness
5	Huyèn	curve	Falling	À	Laxness, breathiness
6	Nạng	drop	Dropping	Ạ	Glottalization/ tenseness

Since it was impossible to create all desired combinations of tones with the same sequence of syllables, almost all of the utterances were of the ‘nonsense’ type. The segmental structure of the utterance was as follows:

nha mai lam nhan nhieu ngo (Vietnamese spelling)

[nja][mai][lam][njan][njo][ngo] (rough phonetic equivalent)

Table 4 lists mean accent command amplitude and timing for the six tones. The timing is expressed relative to the syllabic duration by $T1_{rel}=(T1-t_{on})/(t_{off}-t_{on})$ and $T2_{rel}=(T2-t_{on})/(t_{off}-t_{on})$, where t_{on} and t_{off} denote the onset and offset time of the syllable, respectively.As can be seen, tones 4 and 5 require tone commands of negative polarity whereas the sixth tone does not receive any command at all due to the fact that it did not exhibit any consistent tonal target, but was mainly marked by glottalization. As can be seen from Figure 7, Tones 3 and 6 are often connected with a voicing irregularity where the vocal folds slacken and *F0* drops very low, often to half or even a quarter of the regular value.

Table 4: Mean tone command amplitude and relative timing for the six tones of Vietnamese.

Tone	*N*	*Aa*	$T1_{rel}$	$T2_{rel}$
1	147	.1815	-.0455	.8763
2	120	.3285	.4014	1.0562
3	119	.4618	.4329	.9624
4	37	-.1995	.2473	.7732
5	145	-.0913	.2762	.8179
6	44	.0000	-	-

In order to examine whether synthesized tones could be reliably identified by native subjects, a perception experiment was conducted using stimuli of three different types:

- Natural recordings from the database
- The same utterances as in 1. resynthesized using Fujisaki parameter yielded in the analysis, but without special markers for tones 3 and 6.
- Utterances resynthesized from a sequence of tone 5 syllables employing the mean values of *Aa* and $T1_{rel}$ and $T2_{rel}$ for each of the tones as given Table 4, and vocal fry markers for tones 3 and 6. The phrase component in all of these cases was kept constant.

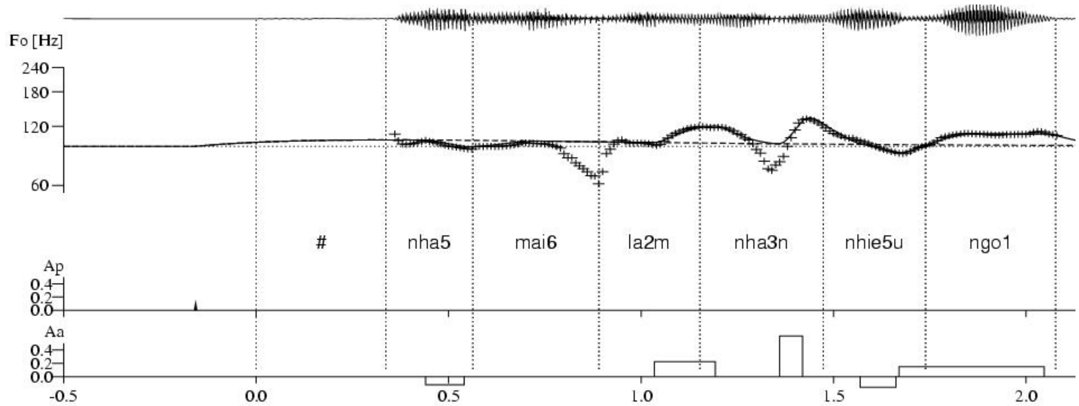


Figure 7: Examples of analysis of two utterances. Top panel: nha5 mai1 la2m nha4n nhie5 ngo1, bottom panel: nha5 mai6 la2m nha3n nhie5 ngo1. The dropping *F0* contours in tones 3 and 6 associated with vocal fry is clearly visible.

20 native speakers of Standard Vietnamese (7 female and 13 male), listened to 60 different stimuli, 20 for each type. Subjects were provided answer sheets with a list containing the written version of the sentence, but without the diacritics indicating the tones, and were presented the stimuli in randomized order. They were asked to complete the diacritics corresponding to the tones they perceived. Table 5 and Table 6 display confusion matrices for the first two types of stimuli. The intended tones and the perceived tones are given in the rows and columns, respectively, along with the total number of judgments *N*. The right-most column lists the percentage of correct judgments. As can be seen, for all three types of stimuli, tones 1, 2, 3 and 5 were correctly identified in well over 90% of judgments whereas tone 4 only yields a correct assessment in about half of the judgments. The drop tone, tone 6, yields relatively poor results in the resynthesized stimuli compared with the natural stimuli. Tone 4 is consistently confused with tone 5 in all three types. These results appear plausible, as tone 4, 5 and 6 are basically all tones with a low onset of *F0*. The results show that the Fujisaki model captures the tonal properties quite well.

Table 5: Confusion matrix, natural stimuli, rows: intended tone, columns: perceived tone. N denotes the total number of judgments, '% corr.' the percentage of correct votes.

T.	1	2	3	4	5	6	*N*	% corr.
1	1305	1	2	3	6	3	1320	98.9
2	6	910	35	0	8	1	960	94.8
3	5	2	662	41	4	6	720	91.9
4	5	4	1	86	51	13	160	53.8
5	34	1	5	23	1370	3	1436	95.4
6	0	0	9	13	7	171	200	85.1

Table 6: Confusion matrix, resynthesized natural stimuli using Fujisaki parameters.

T.	1	2	3	4	5	6	*N*	% corr.
1	1300	2	2	2	9	5	1320	98.5
2	8	907	35	0	7	1	960	94.5
3	9	16	685	0	2	8	720	95.1
4	1	2	4	85	51	17	160	53.1
5	30	1	2	54	1350	3	1436	94.0
6	4	1	6	20	57	112	200	56.0

6. Conclusions

It was shown on examples from four different languages how the Fujisaki model and its extension to a syllable-based integrated model of prosody can be successfully employed for investigating phonological distinctions in the respective languages. The basis for this kind of studies is the establishment of a mapping between prosodic units and structures in a language and the model commands. As we have shown, there is a strong connection between accent commands and accented syllables in German and Finnish, whereas the syllabic tones in Thai and Vietnamese are related to so-called tone commands whose configurations with respect to polarity and alignment need to be carefully selected. Obviously the right choice of commands requires a certain amount of expert knowledge and previous ‘hand-fitting’ of *F0* contours before any automatic algorithm could be developed. Nevertheless, the merits of the approach become clear once the initial ‘barrier’ has been crossed. Considering the fact that even in the ToBI community concern is growing as to the drawbacks of a purely symbolic approach with respect to the modeling of fine alignment of tonal targets with the syllable, as well as the representation of tonal targets of varying ‘height’[19][20], a quantitave approach might be the solution.

LITERATURE

[1] Fujisaki, H.; Hirose, K., “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” *Journal of the Acoustical Society of Japan (E)*, **5**(4), 233-241, 1984.

[2] Fujisaki, H., 1998. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour", in Fujimura, O. (Ed.). *Vocal Physiology: Voice Production, Mechanisms and Functions*. Raven Press Ltd., New York, 347-355.

[3] Pierrehumbert, J., 1980. The phonology and phonetics of English intonation. Ph.D thesis. MIT.

[4] Mixdorff, H., 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. *Proceedings ICASSP 2000*, vol. 3, Istanbul, Turkey, 1281-1284.

[5] Narusawa, S.; Fujisaki, H., Ohno, S., 2000. A method for automatic extraction of parameters of fundamental frequency contours, Proceeedings of ICSLP2000, Beijing, China.

[6] Taylor, P., 1998 The Tilt intonation model. *Proc. ICSLP 98*, vol. 4, 1383-1386, 1998.

[7] Möhler and A. Conkie. Parametric modeling of intonation using vector quantization. *Proceedings of 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

[8] Mixdorff, H. and O. Jokisch (2001): Building An Integrated Prosodic Model of German. In *Proceedings of Eurospeech 2001*, vol. 2, pp. 947-950, *Aa*lborg, Denmark.

[9] A.V. Isačenko and H.J. Schädlich, 1964. *Untersuchungen über die deutsche Satzintonation*. Akademie-Verlag, Berlin.

[10] E. Stock and C. Zacharias, 1982. *Deutsche Satzintonation*. VEB Verlag Enzyklopädie, Leipzig.

[11] Mixdorff, H., Fujisaki, H. (1995): Production and Perception of Statement, Question and Non-Terminal Intonation in German. In: *Proceedings of the ICPhS '95*, Stockholm, Schweden, vol. 2, pages 410-413.

[12] Iivonen, A., 1998. Intonation in Finnish. Daniel Hirst and Albert Di Cristo (eds.) *Intonation systems - A survey of twenty languages*, Cambridge University Press, 311-327.

[13] Mixdorff, H., Vainio, M., Werner, S. and J. Järvikivi (2002): The Manifestation of Linguistic Information in Prosodic Features of Finnish. Proceedings of *Speech Prosody 2002*, pp.511-514, Aix, France.

[14] Vainio, M., Mixdorff, H., Järvikivi, J. and Werner, S. (2003): The production and perception of focus in Finnish. Proceedings of *ICPhS 2003*, Barcelona, Spain.

[15] Luksaneeyanawin, S., “Intonation in Thai,” in Hirst, D. and Di Christo, A. (Ed.), *Intonation Systems. A Survey of Twenty Languages.* Cambridge University Press, Cambridge, 1998.

[16] Mixdorff, H., Luksaneeyanawin, S., Fujisaki, H. and P. Charnvivit (2002): Perception of Tone and Vowel Quantity in Thai. Proceedings of ICSLP 2002, Denver, USA.

[17] Thompson, Laurence. A Vietnamese Reference Grammar. Hawaii: University of Hawaii. 1987.

[18] Mixdorff, H., Hung, N. et al. (2003): Quantitative Analysis and Synthesis of Syllabic Tones in Vietnamese. Proceedings of Eurospeech 2003, Geneva.

[19] Prieto Vives, P. The representation of rising accents in Catalan and Spanish. *Intonation in Language Varieties–the AM approach*, Satellite workshop of ICPhS2003. http://www.vuw.ac.nz/lals/icphs/presentations.htm.

[20] Marotta, G. On the representation of complex Pitch Accents: A new proposal from Tuscan Italian, see[19].