Exploration of Different Types of Intonational Deviations in Foreign-Accented and Synthesized Speech

Matthias Jilka

Department of English Linguistics
University of Stuttgart, Stuttgart, Germany
jilka@ifla.uni-stuttgart.de

10.21437/Interspeech.2005-44

Abstract

The study provides an analysis of the basic manifestations of intonational deviations in foreign-accented (American English accent in German) and synthesized speech. It takes into account the crucial influence of the used model of intonation description and makes a major distinction between individual deviations that cause the impression of foreignness or unnaturalness immediately when they occur, and others that do so only when an accumulation of several such deviations does not allow for a meaningful interpretation anymore. It is argued that this is due to the high variability allowed in prosodic contexts. A closer description of the first group of deviations includes the transfer of categories and of the phonetic realizations of categories as well as a discussion of seemingly unmotivated errors and the most likely causes of intonation errors in synthesized speech. Finally, it is shown that in the case of foreign accent the language-specific manifestations of the presented deviations combine to create a characteristic overall impression of foreignness that is recognizable independently of the segmental content of an utterance.

1. Introduction

It is generally accepted that tonal deviations from the native norm that are relevant for the actual perception of inappropriate intonation are much more difficult to identify than segmental errors.

While an obvious reason is often lack of awareness due to a language-specific preference for segmental perception, there are also a number of linguistically motivated complicating factors. Their identification and classification alone would constitute progress in understanding intonation errors in such different fields as L2 teaching and speech synthesis.

The correct diagnosis of the nature of typically produced and perceived deviations represents an indispensable basis for future improvements in both L2 learners and the prosody generation of speech synthesis systems.

This study examines foreign-accented productions in German by native speakers of American English as well as the intonation generated by the AT&T Natural Voices (R) speech synthesis systems for American English (voice "Crystal") and German (voice "Klara"). All example utterances are available for listening at http://ifla.uni-stuttgart.de/~jilka/devtypes.html

Using a category-oriented system of intonation description (ToBI [1]) it is shown that there are different types of deviations, affecting tonal categories and/or their phonetic realizations, that lead to the impression of non-native or unnatural intonation.

It is claimed that the high variability of form and interpretation inherent in intonation patterns has the consequence that the decision whether a certain tonal configuration is inappropriate strongly depends on the context in which it is uttered.

As a result individual deviations either can be immediately perceived as foreign/unnatural or they simply trigger an alternative interpretation. In this case only an accumulation of such deviations can eventually lead to the perception of foreign-accented or artificial intonation.

Finally, in the case of foreign accent it is shown that the tonal deviations can conspire toward a common effect that is independent of the relationship between intonation and content and that even allows the identification of a language on that basis alone.

2. Influence of the model of intonation description

It must be acknowledged that compared to the identification of segmental errors, intonation presents a problem of representation: the choice of the intonation description/modelling approach influences, even determines the perception of deviations. A description based on the British school model, e.g. [2], obviously expresses foreign-accent differently than a tone-sequence model approach.

Using ToBI not only has the advantage that it is a widely accepted model of intonation description with language-specific inventories, but also that it provides a category-based framework that is compatible with the established, segment-oriented models of foreign accent, e.g., [3], as shown in previous studies such as [4].

The examined utterances and intonation patterns were thus labeled with the respective inventories for American English ([5]) and German (using the Stuttgart ToBI system, [6]). For the analysis of the phonetic realization of equivalent tonal event a parametric approach, PaIntE, ([7]) was chosen which utilizes six parameters, such as steepness of rise, steepness of fall; temporal alignment of peak, amplitude of rise, amplitude of fall and absolute peak height, to determine and measure the shape of an approximation function and thus the F_0 contour in the area associated with a certain tone label.

3. Individual instances of perceived deviations

If a deviating tonal event is unambiguously inappropriate in the given context it is immediately heard as such, i.e. non-native in the case of foreign accent or unnatural in the case of (re)synthesized speech.

3.1. Foreign Accent

3.1.1. Transfer of an L1 category to the L2 within a specific discourse situation

The most easily identifiable cases of intonational foreign accent are those in which a specific discourse situation, as a declarative, a yes/no-question or calling contour etc. is realized inappropriately. A good example is the production of a continuation rise in German by a native speaker of American English.

The continuation rise in German is realized quite simply with a rising nuclear pitch accent spreading to a default boundary tone (L*H % in Stuttgart ToBI notation), whereas in American English there is an additional explicit rise for the boundary tone itself (L+H* L-H%).

Fig. 1 illustrates this in the pronunciation of "lesen können". The middle contour shows the simple rise in the German speaker's version, whereas the American speaker transfers the American English realization of a continuation rise to her German production, leading to an additional fall and rise movement (top contour). The bottom contour depicts a resynthesized version of the American speaker's utterance in which a German intonation pattern has been forced on it (via the tone labels) to correct and make clear the intonation error.

3.1.2. Transfer of an equivalent tonal category with a different phonetic realization

Deviating phonetic realizations of equivalent tonal categories can also be heard as non-native without necessarily being perceived as reflecting an altogether different tonal event. The representation of relevant characteristics is dependent on the descriptive features chosen. An analysis of variance of rising (L*H) pitch accents in German [4] as produced by American and German speakers with the aforementioned PaIntE approach shows significant differences for the steepness of the rise ($p = 0.00015$)), the amplitude of the rise ($p = 0.0000011$) and the duration of the rise ($p = 0.0029$). The results are interpreted in such a way that the rises in L*H pitch accents produced by the Americans are steeper than those produced by the Germans because they have a significantly higher amplitude. The Americans' rises are actually longer but this is outweighed by the greater amplitude. The differences in amplitude (and steepness) are not due to the selection of speakers with the American speakers simply having higher voices. Peaks are on average 216.5 Hz (Germans) and 220.8 Hz (Americans), $p = 0.5177$, thus not significantly different, instead the baseline values are significantly lower for the American speakers ($p = 0.0184$), 169.3 Hz vs. 182.4 Hz, implying that they use bigger pitch ranges.

3.1.3. Seemingly unmotivated intonation errors

Many intonation errors committed by L2 learners are not readily attributable to the influence of their native languages. These errors can take any form from the occurrence of an additional pitch accent or the lack of one to the use of deviating categorical or phonetic realizations. The interpretation of such cases remains speculative. It is entirely possible that many are in fact also instances of transfer that are not recognized due their complexity. In some cases it could also be argued that a simplification in the sense of the concept of a "Basic Variety" [8], leading to a reduced prosodic inventory, is taking place.

Another, all-encompassing, explanation would be to assume that when cognitive demands are too high, speakers get confused to such an extent that they mistakenly assign tonal constellations that do or do not exist in their L1 but that are in any case inappropriate in the discourse situation.

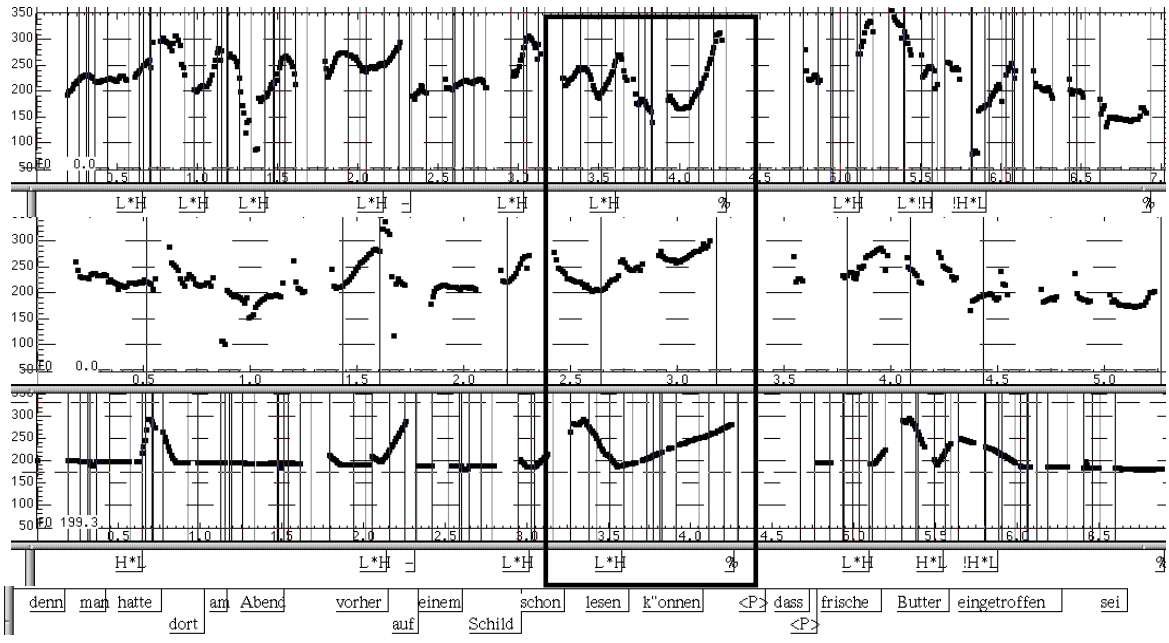

Figure 1. Example for category transfer of a continuation rise on "lesen können". Top contour: American speaker's version with additional fall and rise; middle contour: normal German pattern; bottom contour: rule-generated "improved" version of the American speaker's original

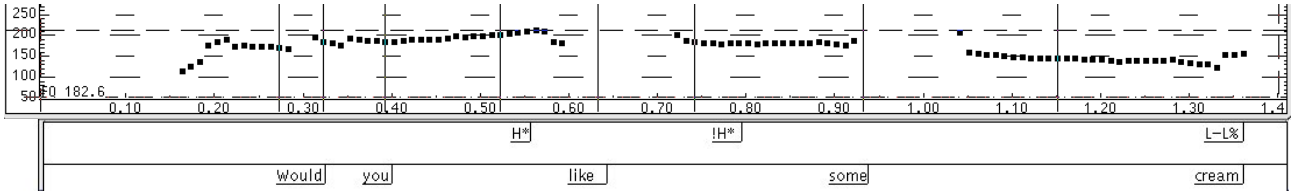

Figure 2. F_0 contour assigned by the AT&T Natural Voices (R) TTS system to the yes/no-question "Would you like some cream?"

3.2. Intonation errors in a text-to-speech system

TTS systems that perform signal processing in order to adjust the synthetic contour to match a predicted F_0 pattern produce intonation errors that are reflections of that particular prediction model. If the model is seen as an independent intonation system atttempting to approximate a target language, then such errors – concerning both category choice and realization - can be considered to abstractly correspond to inappropriate tonal patterns motivated by and transferred from a foreign intonation system.

In TTS systems based on unit selection, on the other hand, the effect of greater weight being attributed to the concatenation of segmental units frequently leads to prosodic aspects being disregarded completely. The example in Fig. 2 shows the text-to-speech synthesis of the yes/no-question "Would you like some cream?" as produced by the AT&T Natural Voices (R) TTS system for American English. The generated declarative-like final fall must be interpreted as clearly inappropriate, even though yes/no-questions ending in falls do occur in natural speech [9].

Synthesized intonation thus often includes rather striking individual intonation errors. Frequently, such errors cannot be labeled satisfactorily with any system of intonation description because they are either completely unnatural or unknown to the language-specific inventory.

4. High variability

Intonation has a high potential for variation, either at random or in dependence of the tonal, segmental or phrasal context.

Phonetic deviations within a tonal category (i.e., from an assumed prototypical realization in the segmental and prosodic context) or the deviating use of whole categories (i.e., their choice and distribution) are, however, not necessarily perceived as foreign-accented or artificial, but may only result in different interpretations as long as the context does not forbid them.

Perception or rather awareness of the deviations is only made possible via a cumulative effect. This is obvious for synthesized speech where repeated intonation patterns, that just by themselves would not be conspicuous, can quickly lead to an impression of unnaturalness.

Similarly, in foreign-accented utterances like the one depicted in Fig. 3 tonal choices that are slightly unusual slowly accumulate from uncommon, but not impossible interpretations to the concluding insight that something is wrong when no sensible interpretation is possible anymore. The first unusually placed tonal category is the rising pitch accent on "auch", here used as a focus particle in the sense of "even". Due to the pitch accent, however, the impression is created that it is used in the more common interpretation of "also" or "too". The following falling accent on "Mehrzahl" reinforces this impression. Similarly, the focus particle "nur" (just, only) is also assigned a pitch accent instead of the following "so" as would be expected. A rising pitch accent is also found on "Zeit" (time), indicating a contrastive focus accent. This L*H pitch accent is followed by yet another L*H accent on "blieb" (stayed) in the very next syllable, creating an uncommonly narrow rise-fall-rise pattern at the phrase boundary. The distribution and high number of pitch accents encourages a forceful interpretation. The listener expects the continuing utterance to describe an extraordinary action or experience of the children described in the preceding context. As the utterance simply continues with a description of how long the children stayed, the listener will unavoidably get the impression that the intonation is inappropriate.

Cumulative effects of this kind are also possible, but less frequent for synthesized speech. The synthesized production in Fig. 4 (top contour), for example, only becomes unnatural with the pitch accent on "hören" which is incompatible with the intonation pattern before it.

5. Overall impression of foreign-accented intonation

In foreign-accented speech the cumulative effects described above combine with the many individual events potentially expressing foreign accent to create one common impression, i.e., several intonational features conspire towards an overall intonation characteristic. In the example of American speakers' intonation in German it can be observed (middle and bottom contour in Fig. 4) that the Americans use about twice as many pitch accents as the Germans in the same stretch of read speech and that they make more use of their pitch ranges (see section 3.1.2.).

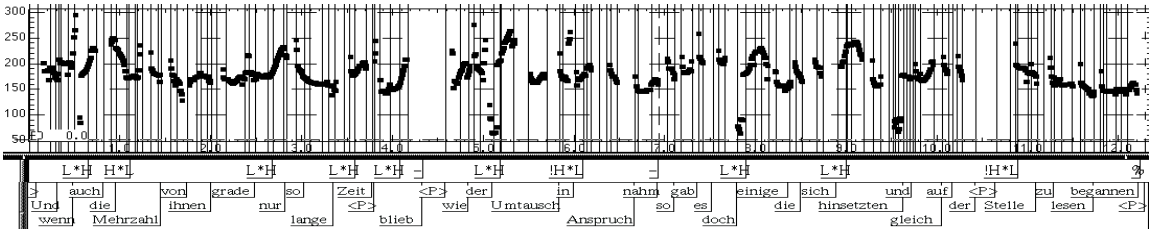

Figure 3. German read speech by American speaker exhibiting cumulative effect of unusual pitch accent choices

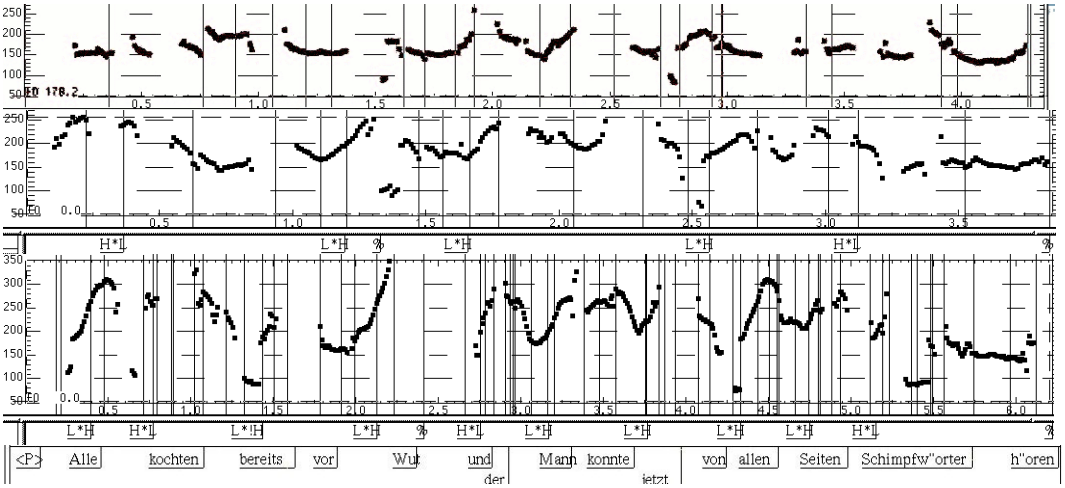

Figure 4. Differences in overall intonation characteristics between productions by German (middle contour) and American speakers (bottom contour). The top contour shows the F_0 contour assigned by AT&T's text-to-speech system for German.

If a transfer of a tonal category takes place, it is likely to lead to additional tonal movements as well, as in the transfer of the continuation rise described in section 3.1.1.

These accumulated patterns create a form of "global" intonational foreign accent that is language-distinctive, if not language-specific, due to L1 influence. It exhibits a certain independence from the segmental level, i.e. the temporal alignment, choice and placement of pitch accents with respect to content/interpretation. This has been demonstrated by a number of studies using various means such as, for example, different stages of delexicalization by means of resynthesis [10, 11]. While many of these studies have shown that rhythmic features alone may be sufficient to distinguish languages, it can also be demonstrated using low pass-filtered speech that rates of language identification accuracy are significantly better when intonation information is involved [4].

6. Conclusion

The present overview offers an analysis of the basic types of intonational deviation in both foreign-accented and (re)synthesized speech. It attempts to heighten awareness and thus facilitate error diagnosis in second language teaching and intonation models used in synthesized speech. The presented results are of course specific to American accent in German and the text-to-speech system examined as well as being dependent on the framework of intonation description used. Nevertheless they indicate that rather general aspects which contribute to the overall impression of inappropriate intonation, such as the number of pitch accents, the pitch range used and category transfer in clearly defined discourse situations could have a relatively large impact in L2 teaching and intonation modelling, if more attention were paid to them.

These characteristics cover a large portion of the foreign- or unnatural-sounding tonal realizations such that improvements in these areas could arguably be equated with major improvements of the overall impression of accentedness or naturalness.

A further advantage is the lack of necessity of a detailed understanding of the contexts that determine the distribution of tonal categories and their respective realizations, as neither linguistically naïve L2 learners nor algorithms of prosody assignment ususually exhibit such an understanding. Resynthesized "correct" versions, as demonstrated in Figures 2 and 4, should provide helpful demonstrative effects as well.

7. References

[1] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J., "ToBI: A standard for labelling English prosody", *ICSLP Proc.*, 867-870, 1992

[2] O'Connor, J. and Arnold, G., *Intonation of Colloquial English*, Longmans, London, 1973

[3] Flege, J. "Second Language Speech Learning: Theory, findings and Problems", in Strange, W. (Ed.) *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, York Press, Timonium, MD, 233-277, 1995

[4] Jilka, M., *The Contribution of Intonation to the Perception of Foreign Accent*. PhD Dissertation. AIMS 6(3), University of Stuttgart, 2000

[5] Beckman, M. and Hirschberg, J., *The ToBI Annotation Conventions*, Ohio State University, 1994

[6] Mayer, J., *Transcription of German Intonation – The Stuttgart System*, Technical Report, University of Stuttgart, 1995

[7] Möhler, G. and Conkie, A., "Parametric Modelling of Intonation Using Vector Quantization", *Eurospeech Proc.*, 1019-1022, 1995

[8] Klein, W. and Perdue, C., "The Basic Variety (or: Couldn't Natural Languages be much Simpler?)", *Second Language Research 13(4)*, 301-347, 1997

[9] Syrdal, A. and Jilka, M., "Exploration of Question Intonation in Read American English", *J. Acoust. Soc. Amer*., *Vol. 115, 2003, p 2543(A).*

[10] Ramus, F. and Mehler, J., "Language Identification with Suprasegmental Cues: A Study Based on Speech Resynthesis", *JASA*, *Vol. 105(1)*, 512-521, 1999

[11] Frota, S., Vigário, M. and Martins, F. "Language Discrimination and Rhythm Classes: Evidence from Portuguese", *Proc. Speech Prosody*, 315-318, 2002