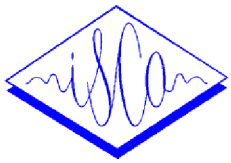
ISCA Archive
http://www.isca-speech.org/archive



Speech Prosody 2006
Dresden, Germany
May 2-5, 2006

The stylization of intonation contours

Grażyna Demenko & Agnieszka Wagner

Institute of Linguistics
Adam Mickiewicz University, Poznań
{lin; wagner}@amu.edu.pl

Abstract

This paper presents the stylization of intonation contours and clustering of F0 movements on accented and post-accented syllables based on annotated speech corpora. Special software – *PitchLine* – has been developed to enable the flexible quasi-automatic segmentation and parametrization of intonation curves. The experimental material obtained from a 15 min passage read by a male speaker included more than 1200 annotated accents and several hundred phrase boundaries. The accuracy of the stylization method was evaluated by measuring the NMSE between original and stylized F0 contours and in a perception study. Stylized F0 contours which were perceived as very different from the original ones required further analysis and re-stylization. Finally, 640 mono-tonal accents formed 6 clusters and 580 bi-tonal accents formed another 6 clusters. The results of clustering confirmed the correctness of the stylization rules.

10.21437/SpeechProsody.2006-48

1. Introduction

The modeling of natural expressive speech features in synthesis and recognition systems requires a representative description of prosody and especially intonation. At present, large acoustic databases of several dozen hours are used in these applications. So far, the annotation of suprasegmental features in speech corpora is rather poor and concerns only elementary prosodic phenomena such as phrase boundaries and lexical accent placement. Intonation models such as ToBI, the Fujisaki Model, Tilt and INTSINT are currently being verified and special attention is being paid to their effectiveness in speech technology applications. The observed lack of synthetic speech naturalness and expressiveness signals the need for using more prosodic information [9], [11]. In speech recognition systems, prosodic features have not yet been used explicitly, even though the preliminary results of prosodic feature implementation in ASR systems showed the possibility of significant error recognition reduction [10]. Thus in various speech technology applications an increasing requirement for an adequate prosody modeling can be observed.

The current study presents the preliminary rules of intonation stylization on the basis of manual annotation of accents and phrase boundaries. In acoustic analysis the goal was to obtain prototypical intonation contours occurring in expressive speech. The goal of perceptual evaluation of stylized F0 contours was the verification of the stylization rules. The approximation rules presented in the Tilt model [13] seem very attractive for the purposes of our study.

The paper is organized as follows: in section 2 the stylization method is discussed. Features and preparation of the experimental material are described in section 3. The next section presents an evaluation of stylization. It is followed by description of F0 contours clustering and its results. The paper ends with a brief discussion on the presented stylization method and annotation of intonation events.

2. Method

2.1. Assumptions

In the development of the rules of the intonation contours stylization and the definition of the fundamentals of the automatic annotation of accents we assume some rules as follows:

- According to the Tilt intonation model [13] an intonation phrase is a sequence of intonation events: R (rises), F (falls) and C (connections), realized by one or more curves.
- Unlike the Tilt model we assign intonation events – rises and falls – to post-accented syllables as well, but only if a post-accented syllable forms an accent group with the preceding accented syllable, which is determined by syntactic context.
- Rs and Fs occur on accented syllables, and Cs occur on unaccented syllables. Rises are realized by rising pitch with a positive slope. Falls are realized by falling pitch with a negative slope. Connections can by realized by either.
- We introduced the AC (level) intonation event. It is associated only with accented syllables and realized by flat pitch with a near zero slope. The treatment of these level accents, on which no F0 change on the syllable can be observed but which acoustic feature give the perception of accentuation, needs further investigation.
- Each curve is connected with the adjacent curve by a straight line.
- The stylization method assumes a specific structure of an intonation phrase:

a) anacrusis or pre-head, labeled ">" (optional)
b) accented syllable or head (we do not distinguish yet between pre-nuclear and nuclear accents), labeled "A" (obligatory)
c) the post-accented syllable, labeled "PA" (optional); forms an accent group with the accented syllable
d) the tail – sequence of unaccented syllables stretching from PA syllable to the phrase end is labeled "=" (optional)

- The F0 stylization quality mainly depends on the accuracy of the F0 contour approximation on accented and post-accented syllables.

2.2. Algorithm

A special software program *PitchLine* was created to perform f0 modeling. It was built in Borland C++ Builder environment on Win32 platform. It required that the input data was segmentated with the *Creatseg* [12] software.
The duration of each segment can be verified manually to allow for the best approximation. The approximation is performed in a quasi-automatic way and the results are shown on screen in real time. Each intonation event is described by the following set of parameters: slope, Fp (F0 value at the start of an event), range of an F0 change and a shape coefficient of the curve. Curves representing intonation events are given by the function defined as follows:

$$\begin{aligned} 0<x<1 \qquad & y = x^{\gamma} \\ 1<x<2 \qquad & y = 2-(2-x)^{\gamma} \end{aligned} \tag{1}$$

The NMSE error is obtained from the following function [1]:

$$NMSE = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}[f(x,y)-f'(x,y)]^2}{\sum_{x=1}^{M}\sum_{y=1}^{N}[f(x,y)]^2} \tag{2}$$

where f is the approximate value function and f` is the approximated value function.
The results of modeling are stored in *.par* and *.f0* files. Files *.par* include parameters describing intonation events and *.f0* includes F0 of the stylized F0 contour in the format of *PitchTier* in *Praat*. This makes it possible to test the quality of modeling using additional software.

3. Experiment

3.1. Material

The material consisted of a 15 minute passage of a chapter from "The Master and Margaret", a novel by M. Bulhakow, which included various sentence modes and a large amount of dialogue. The text was read by a male speaker, who was instructed to read in a moderate tempo and to convey the emotional load of the text. The recordings were done in a professional radio studio.
The recorded material was divided into short passages including from 1 to 4 intonation phrases. As a result 400 wave files containing ca. 1000 phrases were obtained.
Since information about syntactic structure of phrases was needed for the determination of phrase boundaries and the suprasegmental structure of phrases (see subsection 2.1), the whole text was parsed with the PolEng parser [8].
Each phrase was labeled at three tiers: *.lab* (including phonetic segmentation), *.syl* (with segmentation into syllables and intonation phrase constituents) and *.break* (including break indices). Phonetic segmentation was done automatically using the program *CreatSeg* and verified manually. Phonetic transcription was done in the *Polphone* system [3]. Syllable boundaries were determined according to the maximal onset rule and introduced manually on the *.syl* tier. As the purpose of present research is the classification of F0 movements on accented and post-accented syllables, we needed information about accent placement. A group of 5 experienced listeners listened to phrases and marked accented syllables in the text. Each syllable indicated as accented at least by three listeners was labeled "A"(accented) and the following syllable received "PA" label (post-accented).
On the *.break* tier, two types of boundaries were marked: 3 and 4, for minor and major intonation phrase boundaries respectively. A minor intonation phrase has to include at least one nuclear accent and its boundary coincides with the boundary of a syntactic group. A major intonation phrase includes at least one minor intonation phrase and its boundary coincides with a sentence boundary. The linguistic labeling was done in *Wavesurfer*.

3.2. Pitch extraction and stylization

Pitch was extracted every 10 milliseconds using the ESPS method and manually corrected [5]. Faulty F0 values resulting from pitch doubling or halving or laryngelization at the end of phrases were corrected on the basis of a careful auditory analysis. F0 values tracked in unvoiced regions were deleted. Microprosodically affected F0 values occurring at the transitions from unvoiced to voiced parts and vice versa and in the context of voiced plosives and fricatives were corrected only if they caused an inaccurate approximation of F0 curves (see subsection 4.2).
Stylization using *PitchLine* requires a decision as to what kind of function (R, F, C, AC) should be given. This was done by hand marking the element at the start position for each syllable and assigning an appropriate intonation event to each syllable of a phrase according to the minimum of the NMSE.

4. Evaluation of stylization accuracy

The accuracy of the stylization method was evaluated by measuring the NMSE between original and stylized F0 contours and in a perception study. The rationale behind perceptual evaluation was that although the NMSE is an objective method, the correspondence between high NMSE and low perceptual similarity between original and stylized F0 contours is not straightforward. The perception stylization of intonation is not yet known well [4].

4.1. Perception test

Each stylized phrase was resynthesized from *.f0* pitch tier given by the *PitchLine* and original wave file in *Praat.* Five persons participated in the test. They listened to pairs of signals: the original signal and the signal with stylized F0 contour, and assessed the similarity between them.
After listening to a set of 20 original and stylized signals, which enabled identification of types and sources of differences between signals, we decided to use a three point scale for assessment:
a) **1 – identical**
The original and stylized signals are perceived as the same.
b) **2 – a bit different**
Small differences in pitch height (<10Hz) can be perceived between original and stylized signals (e.g. the pitch is too high at the stylized phrase end). They result from erroneous F0 extraction, microprosody or errors in phonetic or syllabic segmentation.
c) **3 – very different**
The pitch of original and stylized signals differs significantly – the signals have different melody. This is caused mainly by unrecognized accents (i.e. a syllable was accented but did not receive "A" label). The other sources of differences are listed in b).The subjects listened to the signals individually, so they could replay the signals if necessary.

4.2. Results

From among 400 stylized phrases, 256 received score 1, 68 score 2 and 76 score 3. After identification of locations, types and sources of differences the stylized signals with the score 3 were stylized again and given for perceptual evaluation once more. Below there is an example of a stylized F0 contour

which initially had a high NMSE and low score in the perception test (i.e. 3: very different from the original sentence). This was caused by: a) the unrecognized accent on dZe (in gentleman), b) microprosodically influenced F0 values at the beginning of the vowel a in tfa- (third accented syllable) and at the beginning of the last accented syllable (brak) of the second phrase, c) wrong location of the last phoneme (k) boundary. Fig. 1 presents the sentence after the first F0 stylization; the re-stylized F0 contour can be seen in the fig.2. It received score 1 in the perception evaluation.

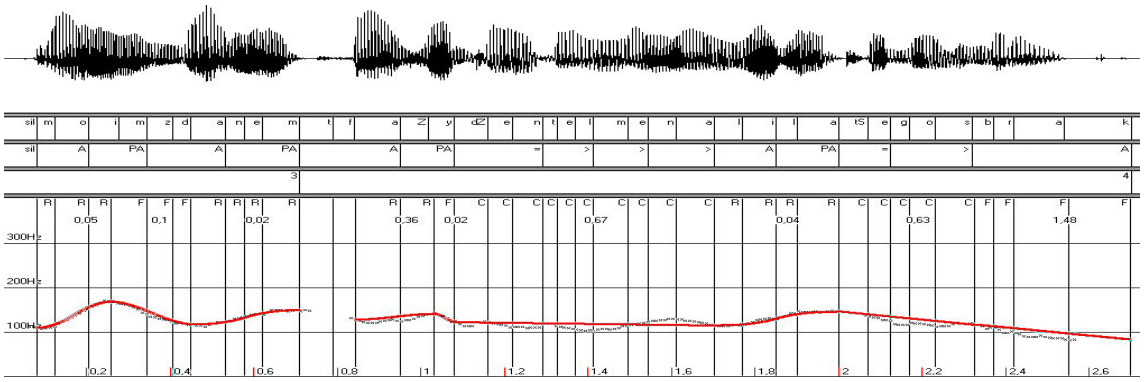


Figure 1: *Sentence:* In my opinion, the face of the lilac gentleman lacks something. *From top to bottom of the picture: waveform, .lab, .syl and .break tiers, and the stylization window. The original F0 contour is marked by dotted black line and the stylized F0 contour in red line.*

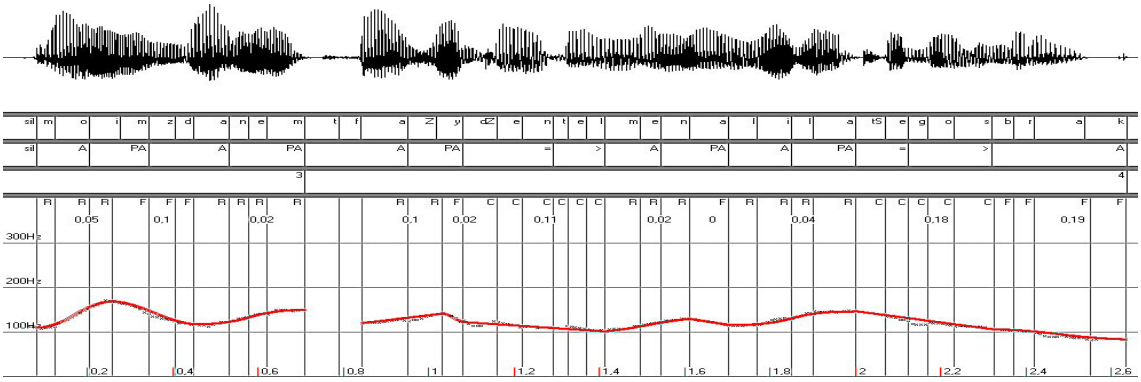


Figure 2: *The same sentence after correction and another F0 contour stylization was perceived as identical (score 1) to the original.*

After a second listening 30% of the re-stylized signals again received the lowest 3 score. These poor results were caused by partial inefficiency of our stylization method. The picture below (fig.3) presents an example: neither F nor C function will approximate the F0 curve on the final accented syllable accurately.

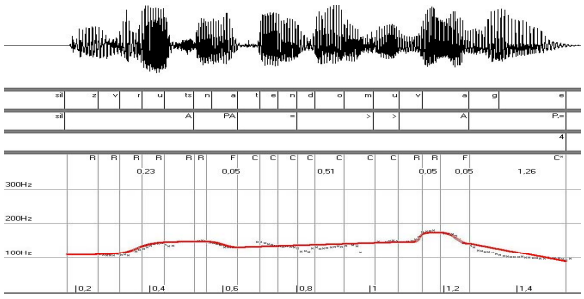


Figure 3: *Sentence:* Have a look at this house. *Due to inaccurate approximation of the final fall the stylized F0 contour was perceived as very different (3) from the original one.*

5. Classification

5.1. General statistics

Preliminary statistics made it possible to estimate the range of variability of intonation event parameters. The table below shows median, minimum and maximum values (in rows) of parameters (in columns) describing accented and post-accented syllables. The last column includes values of the stylization error range.

Accented syllable					
	Slope (Hz/s)	Fp (Hz)	Range (Hz)	bend	error
median	*58,51*	*112*	*14,2*	*1,51*	*0,01*
min	*-357,5*	*70,1*	*-96,8*	*1*	*0*
max	*401,7*	*176,7*	*93,7*	*9,549*	*0,64*
Post-accented syllable					
	Slope (Hz/s)	Fp (Hz)	Range (Hz)	bend	error
median	*-64,5*	*129*	*-13,1*	*1,05*	*0,001*
min	*-529,4*	*65*	*-135,6*	*1*	*0*
max	*364,7*	*208,5*	*86,6*	*9,549*	*0,25*

Table1. *The range of variability of parameters describing accented and post-accented syllables.*

5.2. Clustering

Acoustic analysis of intonation contours allowed for the classification of pitch changes in two groups: mono-tonal pitch changes (rising, falling or level pitch) and bi-tonal

changes formed by different combinations of rising and falling pitch. Mono-tonal accents are based on 640 examples and formed 6 clusters by applying K-means algorithm from S*tatistica* software. Fig. 4 illustrates examples of representative (i.e. the closest to the mean of the cluster) F0 curves. The original F0 is marked with dots and stylized F0 with a line. Classification of F0 curves was based mainly on the direction of F0 change (rise versus fall), range of F0 change and coefficient of the function bend.

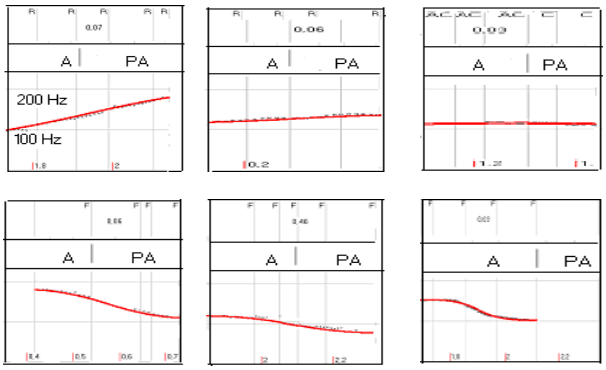


Figure 4: *Mono-tonal accents - examples of representative (i.e. the closest to the mean of the cluster) F0 curves.*

Bi-tonal accents based on 580 examples formed 6 clusters by applying the same K-means algorithm. In the classification, the most frequent combinations of rising and falling F0 curves with different slopes, shapes and ranges were taken into account. Figure 5 presents examples of clusters most representative in a given group.

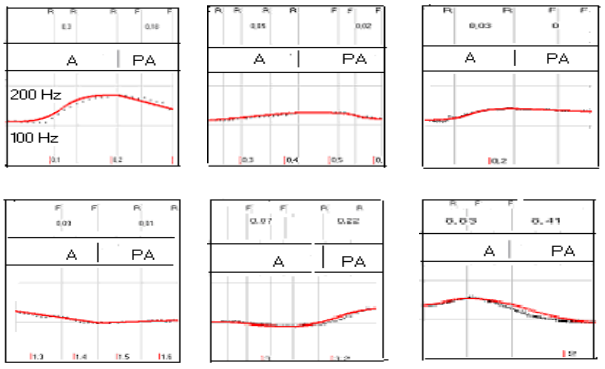


Figure 5: *Bi-tonal accents - examples of representative (i.e. the closest to the mean of the cluster) F0 curves*.

6. Conclusion and discussion

In general, results of the perceptual evaluation of stylized F0 contours proved the correctness of the stylization rules. The clustering results are in line with conclusions of previous research in F0 contour classification in Polish [2], but they are not final. In this study, the annotation of pitch peak location (early versus late peak) has not been taken into consideration, as there were too few examples of early/late peaks and no statistical evaluation could be carried out.

The classification was carried out only for accented and post-accented syllables. But still, fragments of F0 contours both preceding and following accented syllables require detailed analysis. This would make it possible to determine intonation contexts, in which specific F0 changes on accented and post-accented syllables may occur.

So, the results obtained in this study should be treated as preliminary. We assume the possibility of the stylization method optimization through analysis of more expressive speech produced by more speakers. First, we will have to solve the problem of pitch normalization across speakers. Different solutions are indicated in the literature [4], [6]. The necessity of accounting for suprasegmental analysis resulting from voice quality is also indicated [7]. An extremely low F0 within a speaker's range is associated with glottal fry or creakiness. It is likely that the listener can directly use information about voice quality to locate pitch at least at the base or top of a speaker specific range [7]. The definition of the acoustic correlates of accents also requires verification. Especially the annotation of level accents is debatable. Most often, these accents are defined as realized by duration and on the basis of rhythmic conditions. There are also plans to carry out an extensive analysis of acoustic and phonetic features to determine the prosodic structure of sentences. We hope that this will improve our framework for automatic intonation pattern description and recognition.

7. References

[1] Barańczuk, Z., 2000. In Proceedings of Multimedial Information Systems, *Multimedialne i Sieciowe Systemy Informacyjne* (MISSI 2000)

[2] Demenko,G., 1999. Analysis of Polish suprasegmentals for needs of Speech Technology, ed. UAM, Poznań.

[3] Demenko, G.; Wypych, M.; Baranowska, E., 2003. Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis, *Speech and Language Technology*, Poznan, ed. PTFON, vol.7 (79-95)

[4] t'Hart J.; Collier R.; Cohen A., 1990). A perceptual study of intonation, Cambridge University Press, Cambridge.

[5] Hess, W., (1983) Pitch determination of Speech Signals, Springer Verlag, New York.

[6] Hirst, D., 1992. Prediction of Prosody: An overview. In *Talking Machines*, Bailey, G., Benoit, C., ed. North Holland.

[7] Honorof, D.N.; Whalen, H., 2005. Perception of pitch location within a speaker's F0 range, *JASA*, 117 (4), 2193-2200,

[8] Jassem, K., 2002. Transfer w systemie POLENG-3 in: *Speech and Language Technology*., Poznan, ed. PTFON, vol.6

[9] Klabbers, E.; Santen, J., 2004. Clustering of foot-based pitch contours in expressive speech, In Proceedings of *ISCA Speech Synthesis Workshop*, Pittsburgh

[10] Milone, D.H.; Rubio, A.,J., 2003. Prosodic and Accentual Information for Automatic Speech Recognition, *IEEE Transactions on Speech and Audio Processing*, vol.11, no.4.

[11] Santen, J.; Kain, A.; Klabbers, E.; Mishra,T., 2005. Synthesis of prosody using multi-level unit sequences, *Speech Communication 46,* 365-375

[12] Szymański, M.; Grocholewski, S., 2005. Dynamic programming method for fine-tuning the boundary points in automatic segmentation of speech, *Archives of Acoustics,* vol.30, no.3, PAN ed. Warszawa.

[13] Taylor P., 2000. Analysis and synthesis of intonation using the Tilt model, *JASA,* vol.107, no.3

This research has been carried out under grant nr 11C 003827 received from Polish Ministry of Scientific Research and Information Technology.