Automatic Pitch Stylization Enhanced by Top-Down Processing

Mikołaj Wypych

Institute of Fundamental Technological Research
Polish Academy of Sciences, Warsaw, Poland
mik@polfonetika.com

Abstract

In the paper an original method of pitch stylization from the speech waveform and its orthographic transcript is presented. In addition to bottom-up data processing, a top-down step is employed. The top-down step allows for the reduction of contextual variability of intonational structure constituents. Software implementation of the stylization method for the Polish language is described. The design takes advantage of components borrowed from an existing automatic intonation recognizer. Fundamental frequency extraction in the design is performed using a comb filter. In a subsequent stage, a syllable-wise pitch stylization is performed, followed by contextual pitch tracking. Intonational structure is recognized by an intonational parser based on Hidden Markov Models. The intonation model conveying an annotation system is taken from the recent intonation grammar for Polish by Jassem. Components of the design were developed in parallel which allowed for the coordination of tradeoffs between the modules. Training set and exemplary results are presented together with a discussion of future improvements.

1. Introduction

Fundamental frequency (F_0) is the greatest common divisor of the frequencies of the harmonics in an acoustic signal. Pitch is understood in the article as a perceived F_0 of speech. Several pitch representations have been in use in speech technology. Compared to detailed F_0 measurement, pitch representation does not need to include microprosodic variations. It may depend on spectral features of speech and it should not contain unperceived F_0 discontinuities (e.g. octave jumps). We distinguish three levels of abstraction of pitch representations: acoustic, phonetic and phonological.

On the acoustic level, pitch is represented in the form of piecewise continuous function of time. An instance of acoustic pitch representation for an utterance is called a pitch trajectory. The act of obtaining a pitch trajectory is called pitch extraction. The analytical form of pitch trajectory is typically used in speech synthesis systems. In speech analysis systems, sampled representation of pitch trajectory is preferred. Once again, note the difference between acoustic representation of pitch (pitch trajectory) and acoustic representation of F_0 (F_0 trajectory). The former characterizes perception with the latter representing the physical feature of the sound wave. Acoustic pitch representations are language independent and allow for lossless reconstruction of pitch.

Phonetic representation of pitch describes pitch as a sequence of events. Each event stores information on pitch over a limited period of time by means of a vector of continuous variables. Phonetic representation of pitch can be seen as a loosely compressed acoustic representation in which the losses do not influence perception of speech fragment. Both the act of obtaining phonetic representation of pitch and an instance of phonetic pitch representation are called pitch stylization. It is often the case that pitch stylization is obtained from pitch trajectory. Phonetic representation of pitch may be language independent.

Phonological representation of pitch is intended to convey only such aspects of pitch as may differentiate meaning. Especially, unlike the lower-level pitch representations, phonological representation does not carry extra-linguistic content. Phonological representation of pitch is categorial – it consists of finite sequences of events described by limited set of categories. In this article, we will reserve the term intonation for phonological representation of pitch only. The intonation model comprises a set of intonation categories, grammar governing the construction of their sequences and a mapping between sequences of categories and lower-level pitch representation. Several intonation models have been proposed among which ToBI-related and British School-related models seem to be the most influential. Coexisting standards of intonation representation like SAMPROSA, INTSINT and multiple variations of ToBI show that currently, there is no widely accepted universal model for intonation representation. As a result, the selection of an intonation model is an important design choice for a speech engineer.

The act of obtaining a phonological representation of pitch is called intonation parsing (or intonation decoding if we use finite-state intonation grammar). The instantiation of phonological representation is called intonational structure. Depending on design choice, intonation parsing is performed directly on pitch trajectory or on pitch stylization. Phonological models are language specific.

It may be interesting to compare sizes for the pitch representations. A 30 minute subset of our speech corpus requires approx. 720kB for sampled acoustic pitch representation, approx. 88kB for phonetic pitch representation and approx. 2400B for phonological pitch representation.

Pitch stylization has been performed in several ways starting from purely manual approaches to fully automatic ones. Some of the most interesting examples of manual pitch stylization include transcriptions used by Steele [1], IPO perceptual method [2] and “tadpole” stylization by Arnold and O’Connor [3]. Manual pitch stylization is a labor intensive task and introduces interlabeller agreement problems. Some of the well-known examples of automatic pitch stylization methods include: Prosogram [10], Tilt [17] and Momel [6]. Automatic pitch stylizers are sensitive to F_0 extraction errors and can produce quite different results depending on the settings of heuristic parameters.

Typical applications of pitch stylization are: representing pitch in speech synthesis and speech recognition systems, annotation of corpora for corpus-based speech synthesis, intonational research, didactics and logopedics.

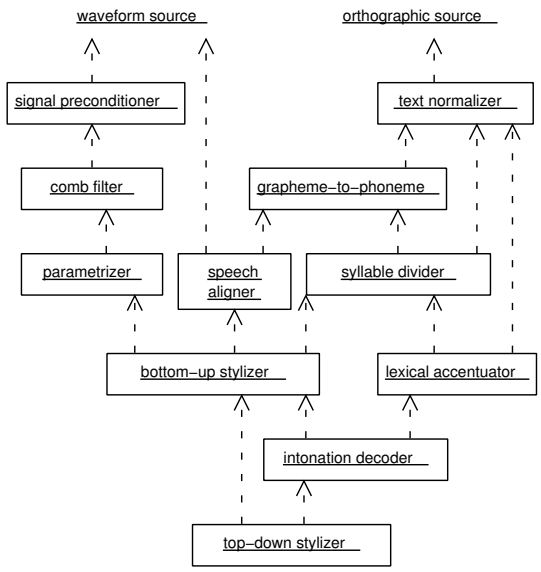

Figure 1: *Components and dependencies in the automatic pitch stylizer*

In this paper we submit an original method of automatic pitch stylization with the use of intonational structure.

2. Inside the automatic pitch stylizer

The present automatic pitch stylizer can be split into four functional blocks: weighted fundamental frequency extractor, bottom-up pitch stylizer, intonation decoder and top-down pitch stylizer. The first three blocks are based on the components of an existing automatic intonation recognizer (see [18] for more details). The input of the present pitch stylizer consists of a speech waveform and its orthographic transcript. Figure 1 shows components comprising the pitch stylizer. More details on the components are provided in the following subsections.

2.1. Weighted fundamental frequency extractor

The weighted fundamental frequency extractor produces a F_0 trajectory (not a pitch trajectory) and an additional weighting trajectory for a given speech waveform. The weighting trajectory is intended to indicate perceptual prominence of F_0 at a given time point. The weighting trajectory is close to or equal to zero for silent, unvoiced sections of speech waveform and achieves high values for loud and highly harmonic sections of speech waveform. Internally, the extractor consists of three components: signal preconditioner, comb filter and parametrizer.

The signal preconditioner filters out intonationally insignificant spectral components from the input waveform and reduces data rate to 64 kbps. It is implemented by means of preemphasis (a=.97), low-pass filtering (cutoff=2kHz), downsampling and DC-removal. The comb filter used in the extractor is a frequency-domain filter with logarithmically located combs. Time-frequency transformation in the comb filter is performed using 256 point FFT with 16 ms step. The coefficients of the comb filter are computed automatically during a self-training routine with the use of a gradient descent algorithm. The parametrizer transforms the comb filter response into two real values. The first represents the location of a maximum response comb. The second represents the difference between the maximum and the minimum response of the combs in the filter.

Please note that the presented F_0 extractor does not apply any pitch-tracking step. Especially the issues accompanying fundamental frequency extraction (e.g. vocal-fry, devoicing, F_1 interference) are left to be resolved by the higher-level better informed processing blocks.

Compared to F_0 extraction methods based on autocorrelation, comb-filtering makes better use of the second and the higher harmonics of the speech signal. Given the flexibility of our implementation, setting a tradeoff between F_0-division and F_0-multiplication errors was possible. Our extractor is tuned up to minimize the number of F_0-multiplication errors. The resulting percentage of pitch multiplication errors in our speech corpus is equal to 0.327% measured syllable-wise (the error is detected if at least one of extraction point in a syllable is multiplied).

High processing performance of the extractor was achieved thanks to the Integrated Performance Primitives (IPP) library for signal processing by Intel. The IPP library takes advantage of SIMD vector processing instructions supported by most modern PC CPUs.

2.2. Bottom-up pitch stylizer

The input data of the bottom-up pitch stylizer are: acoustic F_0 representation, speech waveform and orthographic transcript. Speech waveform and orthographic transcript are necessary for the accurate detection of syllable boundaries in the acoustic F_0 representation.

The input orthographic transcript is partially normalized and chopped into letters. The resulting sequence of letters is passed through grapheme-to-phoneme (g2p) component and syllable divider. The employed g2p component is based on a improved rules from [8]. (One of the earlier incarnations of the component was presented in [7].) The syllable divider is based on finite state syllable division rules described in [9]. Additionally to syllable division rules, that work on a sequence of phonemes, a set of rules working on a sequence of a orthographic letters was added in order to support morphologically/orthographically based syllabification exceptions. The rules and a dictionary of exceptions are compiled into Finite State Automata using a specialized compiler incorporating FSA6 toolkit ([5]).

The bottom-up stylizer performs forced alignment of syllable boundaries using the Sonic speech recognition engine ([11]). The Sonic was trained for Polish in the project described in [12].

After the syllable locations in speech are known, the stylizer processes the F_0 trajectory syllable-wise. The phonetic representation used in the system is original and is based on what we call *stylization segments*. The stylization segment is a real vector describing pitch over a part of a syllable. Two stylization segments are assigned to each syllable. The first stylization segment is associated with the beginning-to-middle part of the syllable and the second is assigned to the middle-to-end part of the syllable. The definition of the middle point of the syllable is crucial for the algorithm and as was found, the best-performing solution is based on the center of mass of weighting trajectory of the syllable. Elements of the stylization segment are: pitch

height, glissando and duration.

Height and glissando parameters are computed from the F_0 trajectory with the use of the weighting trajectory. As a result loud, highly harmonic fragments of the syllable nucleus have stronger effect on the parametrized value than the other parts. Note that according to a long analysis window used in the F_0 extractor, there is no need for to deal with short-term F_0 extraction issues (e.g. jitter or microprosodic bursts).

After the syllable-wise stage of stylization, a pitch tracking stage occurs. The pitch tracking stage is intended to reduce the number of fundamental frequency extraction errors and eliminate pitch halving caused by vocal fry. The algorithm works on a window of at most 7 stylization segments and takes advantage of the feature of our comb-filtering extractor mentioned above: if it fails in pitch determination it is much more likely to divide the pitch value than to multiply it. In general, the tracking algorithm maximizes the mean pitch height for the window, simultaneously minimizing the sum of the absolute values of pitch differences between succesive stylization segments.

The method revealed high accuracy (99.4% measured syllable-wise) for our testing set comprising male and female voices.

2.3. Intonation decoder

The intonation decoder recognizes the intonational structure of speech using intermediate speech and text representations available in the system (see 1 for detailed dependencies). The intonation model used in the decoder is the British School style model proposed for Polish by Jassem ([4]). The decoder uses phoneme symbols string and syllable boundaries in the string for lexical accent assignment. Lexical accentuation rules used in the system were specialized for the intonation recognition task.

The intonation decoder is based on the Continuous Density Gaussian Mixture Hidden Markov Model (HMM). Preparation of the observation sequence for HMM starts with augmenting the pitch stylization with the lexical accent parameter. Additionally, delta and delta delta coefficients are calculated. Selected dimensions of the resulting vectors become an observation sequence for the HMM (see [18] for more details).

In the current version of the decoder the input stream of HMM consists of 4 dimensions whose pdfs are modeled by 2-component Gaussian Mixtures in each state. Extensive parameter tying is employed in order to avoid data sparsity problem.

The topology of the HMM is based on finite state intonation grammar accompanying the intonation model. The intonation grammar defines the intonational phrase as [wPT]{sPT}NT (using BNF notation), where: wPT is a weak prenuclear tune – a sequence of syllables without real accent and sPT/NT are strong prenuclear and nuclear tunes – sequences of syllables starting at a syllable bearing real accent. The finite-state character of the grammar and the structural approach to intonation analysis (as opposed to sequential intonation models like ToBI) makes it feasible for HMM modeling. The Jassem model of intonation specifies several instances of tunes in each of the classes. Respective fine-grained probabilistic finite-state grammar is inferred from the training corpus.

The decoder was implemented by means of HTK3 toolkit ([13]).

2.4. Top-down pitch stylizer

The functional blocks described so far, process data in a bottom-up manner i.e. they derive more abstract representations from less abstract ones. The top-down pitch stylizer works in an opposite direction refining existing pitch stylization by taking advantage of the intonational structure. The goal of the top-down pitch stylizer in our method is to reduce contextual variability in the pitch stylization. The contextual variability of pitch stylization is understood here as regular alternations of stylization segments in tunes determined by contextual factors. Major contextual factors are: the categories of surrounding tunes, the segmental content of the tune and the pattern of lexical accents in the tune.

On reanalysing pitch stylization of our speech corpus we selected two exemplifications of contextual variability for processing within the top-down stylizer:

1. the rise of the penultimate stylization segment in falling sPTs that are followed by tunes starting with a higher pitch,
2. the lowering of the initial stylization segment in falling sPTs that are preceded by tunes ending with a a lower pitch.

Contextual variability arising from the phenomena is cancelled out using the following tentative algorithm. Let S[0], S[1], ..., S[n-1] represent the sequence of stylization segments in falling sPT.

1. **if** S[n-2].pitch $>$ S[n-3].pitch **then** S[n-1].pitch $=$ S[n-2].pitch $=$ S[n-3].pitch, S[n-1].glissando $=$ S[n-2].glissango $= 0$.
2. **if** S[0].pitch $<$ S[1].pitch **then** S[0].pitch $=$ S[1].pitch, S[0].glissando $= 0$.

3. Software integration

The present stylizer was implemented using engineering techniques of different genres. The stylizer contains speech processing and text processing techniques, data-based and rule-based techniques, bottom-up and top-down techniques. Data flow on the component level shows that the pipeline architecture may be inefficient in the system implementation (see figure 1). Thus, a specialized Software Architecture for Language Engineering (SALE) was borrowed from the automatic intonational recognizer presented in [18] (see [14] for more on SALEs). Our SALE implements a data-driven software integration approach resembling Blackboard architecture. Components in the architecture do not communicate directly but, rather use a shared data structure. All the changes made to the structure are broadcast through interested components. The shared data structure in our system is based on a modified version of the Annotation Graph described in [15].

4. The training corpus

The training corpus for the present stylizer consists of a subset of spontaneous Polish speech from PoInt corpus described in [16]. The original intonation labeling of the training corpus was provided by Jassem. In the subsequent stage, the labeling was verified and agreed among other labelers. Additional data as speech waveform anchoring and metadata was added to the training corpus. The resulting corpus consists of 30 minutes of speech from male and female subjects. There are 468 intonational phrases, 1327 tunes and 3697 syllables. The training corpus is used for training HMMs coefficients in the intonation decoder and for inferring fine-grained transition probabilities in the intonational grammar.

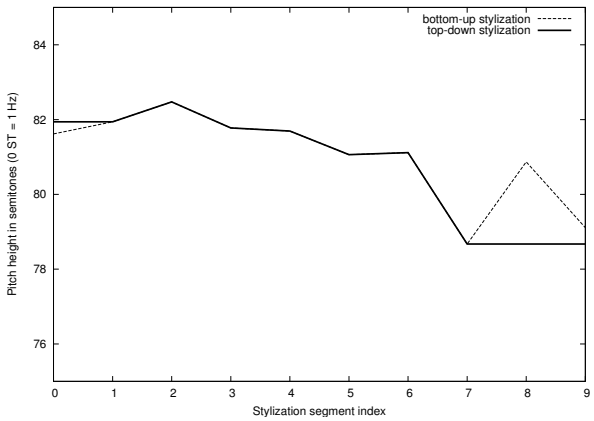

Figure 2: *Stylization segments of sPT "nie.pa.mię.tam.tych" from a stentence "Nie pamiętam tych imion." (I do not remember these names.) by male voice.*

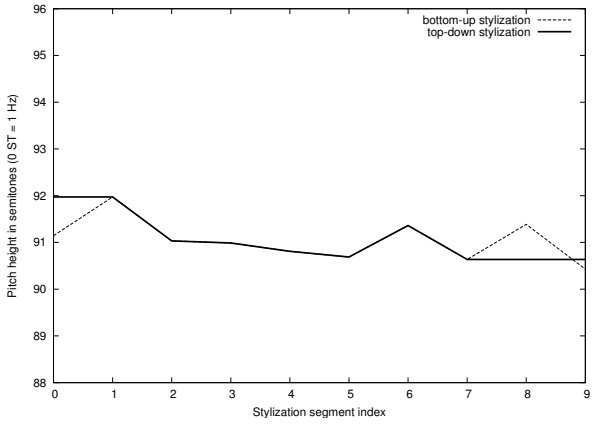

Figure 3: *Stylization segments of sPT "dzie.ję.że.tak.to" from a stentence "Mam nadzieję, że tak to potrwa." (I hope it will last.) by female voice.*

5. Exemplary results

Figures 2 and 3 show the effects of applying top-down refinements to bottom-up stylization on examplary strong prenuclear tunes from our speech corpus. The dashed line represents bottom-up stylization and the solid line represents the stylization after the top-down processing step.

6. Conclusion and future work

In this paper, we presented a pitch stylizer that uses the intonational structure of an utterance for the refinement of its pitch stylization. We demonstrated the usefulness of the approach in the removal of selected contextual influences between stylizations of intonational phrase constituents (tunes). The method of stylization may be useful for speech analysis systems because it reduces contextually determined variability. The method also may be advantageous for pitch stylization in didactics because it filters out some linguistically insignificant components from the stylization that could distract the student. Our future work includes the extension of the top-down stage to a greater number of tunes and contextual phenomena, possibly with the use of trainable methods. Finally, we are also interested in comparing perceived differences (if any) of resynthesized pitch stylizations with and without top-down processing.

7. References

[1] Gussenhoven, C., The Phonology of Tone and Intonation, Cambridge University Press, Cambridge, 2004.

[2] 't Hart, J., Collier, R., Cohen, A., A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody, Cambridge University Press, Cambridge, 1991.

[3] O'Connor, J. D., Arnold G. F., Intonation of Colloquial English, Longman, London, 1973.

[4] Jassem, W., Real and Potential Accent in Spontaneous Polish, (in preparation).

[5] van Noord, G., FSA 6 toolkit, http://odur.let.rug.nl/˜vannoord/Fsa/, 2004.

[6] Hirst, D., Espesser, R., "Automatic Modelling of Fundamental Frequency using a quadric spline function", in Travaux de l'Institut de Phonetique d'Aixen-Provence, 15, 75-85, 1993.

[7] Wypych, M., Demenko, G. and Baranowska, E., "A Grapheme-to-Phoneme Transcription Algorithm Based on the SAMPA Alphabet Extension for The Polish Language", in the Proceedings of International Congress of Phonetic Science, Barcelona, 2003.

[8] Steffen-Batog, M. and Nowakowski, P., "An Algorithm for Phonetic Transcription of Orthographic Texts in Polish", Studia Phonetica Posnaniensia, vol. 3, Poznań, 1992.

[9] Jassem, W., Syllable division rules for Polish (unpublished work).

[10] Mertens, P., "The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Model", in the Proceedings of Speech Prosody 2004, Nara, 2004.

[11] Pellom, B. and Hacioglu, K. Sonic: The University of Colorado Continuous Speech Recognizer, Technical Report, CSLR, University of Colorado, Boulder, 2004.

[12] Dziubalska, K., Cole, R., Pellom, B., Sobkowiak, W., Wypych, M., Bogacka, A., Ma, J., Struemph, T., Krynicki, G., "The Use of Metalinguistic Knowledge in a Polish Literacy Tutor", in Proceedings of GlobE, Warsaw, 2004.

[13] Young., S., Kershaw, D., et. al., The HTK Book, Microsoft Corporation, 2004.

[14] Cunningham, H., Software Architecture for Language Engineering, PhD thesis, University of Sheffield, 2000.

[15] Bird, S. and Liberman, M., "A formal framework for linguistic annotation", Speech Communication 33 (1,2), 2001.

[16] Karpiński, M., "The Corpus of Polish Intonational Database (PoInt)", Investigationes Linguisticae VIII, Poznań, 2002.

[17] Taylor, P. A, "The Tilt intonation model", in Proceedings of ICSLP98, Sydney, 1998.

[18] Wypych, M., "An Automatic Intonation Recognizer for the Polish Language Based on Machine Learning and Expert Knowledge", in Proceedings of Interspeech 2005, Lisboa, 2005.