Wavelets for intonation modeling in HMM speech synthesis

Antti Suni[1]*, Daniel Aalto*[1]*, Tuomo Raitio*[2]*, Paavo Alku*[2]*, and Martti Vainio* [1]

Institute of Behavioural Sciences (SigMe Group), University of Helsinki, Finland
[2]Department of Signal Processing and Acoustics, Aalto University, Finland
antti.suni@helsinki.fi, daniel.aalto@helsinki.fi, tuomo.raitio@aalto.fi
paavo.alku@aalto.fi, martti.vainio@helsinki.fi

Abstract

The pitch contour in speech contains information about different linguistic units at several distinct temporal scales. At the finest level, the microprosodic cues are purely segmental in nature, whereas in the coarser time scales, lexical tones, word accents, and phrase accents appear with both linguistic and paralinguistic functions. Consequently, the pitch movements happen on different temporal scales: the segmental perturbations are faster than typical pitch accents and so forth. In HMM-based speech synthesis paradigm, slower intonation patterns are not easy to model. The statistical procedure of decision tree clustering highlights instances that are more common, resulting in good reproduction of microprosody and declination, but with less variation on word and phrase level compared to human speech. Here we present a system that uses wavelets to decompose the pitch contour into five temporal scales ranging from microprosody to the utterance level. Each component is then individually trained within HMM framework and used in a superpositional manner at the synthesis stage. The resulting system is compared to a baseline where only one decision tree is trained to generate the pitch contour.

Index Terms: HMM-based synthesis, intonation modeling, wavelet decomposition

1. Introduction

The fundamental frequency (f_0) contour of speech contains information about different linguistic units at several distinct temporal scales. Likewise prosody in general, f_0 is inherently hierarchical in nature. The hierarchy can be viewed in phonetic terms as ranging from segmental perturbation (i.e., microprosody) to a levels that signal phrasal structure and beyond (e.g., utterance level downtrends). In between there are levels that signal relations between syllables and words (e.g., tones and pitch accents). Consequently, the pitch movements happen on different temporal scales: the segmental perturbations are faster than typical pitch accents, which are faster than phrasal movements and so on. These temporal scales range between several magnitudes from a few milliseconds to several seconds and beyond.

In HMM-based speech synthesis paradigm, all modeling is based on phone sized units. In principle, slower intonation patterns are more difficult to model than segmentally determined ones. Moreover, the statistical procedure of decision tree clustering highlights instances that are more common, resulting in a good reproduction of microprosody and overall trends (such as general downtrends) and relatively poor reproduction of prosody at the level of words and phrases. This shortcoming calls for methods that take into account the inherent hierarchical nature of prosody.

Traditionally the problem has been approached by using superpositional models which separate syllable and word level accents from phrases [2, 7]. On feature extraction side, discrete cosine transform parameterization of f_0 has been investigated, providing compact representation of the pitch contour [12]. Typically, each voiced segment or syllable and phrase are parameterized with a constant number of DCT coefficients, statistical clustering is performed based on contextual features, and synthesis is performed in additive fashion [11]. However, the constant number of coefficients is problematic for variable length units, and natural continuity between units is difficult to achieve.

In HMM framework, decomposition of f_0 to its hierarchical components during acoustic modeling has been investigated [4, 15]. These approaches rely on exposing the training data to a level-dependent subset of questions for separating the layers of the prosody hierarchy. The layers can then be modeled separately as individual streams [4], or jointly with adaptive training methods [15]. Results indicate that syllable level modeling improves prosody whereas higher levels do not provide benefits.

In HMM-based speech synthesis, f_0 is modeled jointly with voicing decision. The unit of modeling is typically a phone HMM with five states. For each state, predefined contextual questions concerning phones, syllables, words and phrases are used to form a set of possible splits in a decision tree. The splitting decisions are made in a greedy fashion based on likelihood increase. Thus the hierarchical nature of intonation is only implicitly addressed by questions on different levels of hierarchy. With multiple levels, including voicing decision, modeled by a single set of trees, the rare or slow events can not be modeled robustly, due to fragmentation of the training data by previous, more urgent splits for the short time scale of the model.

In this paper, we present a solution to the problems outlined above based on continuous wavelet transform (CWT). The CWT is used to decompose the f_0 contour into several temporal scales that can be used to model the levels ranging from microprosody to the utterance level separately. As well as separating the contour into meaningful temporally assigned levels – ranging from microprosody to utterance level prosody – the CWT produces a continous f_0 contour which has further merits. Earlier, wavelets have been used in speech synthesis context for parameter estimation [3, 6, 10].

We chose four f_0 modeling methods for comparison: (1) The normal HTS method using the MSD stream, and two

wavelet-based setups modeling the f_0 contour on several distinct levels: (2) one with a joint model and (3) one where five separate CWT based levels are modeled separately. In addition, (4) a continous interpolated f_0 stream model was added. The fourth method was added in order to evaluate the wavelet based methods against another model using continuous trajectories since interpolation alone has been reported to improve f_0 modeling [14].

Objective comparison of the proposed methods is presented against single-stream baselines using two GlottHMM [9] Finnish voices trained from a male and a female corpus.

2. Pitch decomposition and wavelets

2.1. Extraction and preprocessing of f_0

GlottHMM vocoder was used for estimating the fundamental frequency (f_0) of speech. GlottHMM is a physiologically oriented vocoder that uses glottal inverse filtering for separating speech into the glottal source signal and the vocal tract filter. The iterative adaptive inverse filtering (IAIF) method is used for the separation, and the f_0 is estimated from the glottal source signal that is free from the distracting vocal tract resonances [9].

The autocorrelation method [8] was used to estimate the f_0. A range of possible f_0 values is defined based on the speaker's f_0 range in order to reduce gross errors. The voiced-unvoiced decision is made based on the energy of the low frequency band (0–1 kHz) and the number of zero-crossings in the frame. The length of the frame from which the f_0 is estimated is longer than the speech analysis frame in order to estimate the lowest possible f_0 values, as low as 30 Hz. The frames determined as unvoiced are marked as zeros. Parabolic interpolation was used in order to reduce the estimation error due to finite sampling period; a quadratic function is fitted to the peak of the autocorrelation function (ACF) to find the refined f_0 value.

Finally, post-processing is applied to the estimated f_0 trajectory. A repetitive process is applied which consists of 3-point median filtering, filling small unvoiced gaps and removing outlier voiced sections, and detection of unnatural discontinuities based on weighted linear estimation of each individual f_0 estimate from previous and following samples. If the difference between the estimated and the actual values is greater than a specific threshold (based on the mean and variance of the f_0 trajectory), the original value may be replaced with a secondary f_0 estimate from the ACF. This replacement depends on the goodness of the fitting and the relative jump of the original f_0 estimate. An example of extracted f_0 is shown in the top pane of Figure 1.

2.2. Completion of f_0 over unvoiced passages

The wavelet method is sensitive to the gaps in the f_0 contour and therefore, the f_0 contour is completed to yield a continuous f_0 trajectory. Since the wavelet approach aims at connecting the signal to the perceptually relevant information, the linear frequency scale is transformed to the logarithmic semitone scale. A simple linear interpolation method is used. First, smoothed version of the original f_0 was created, and then interpolated over unvoiced passages. The smoothed unvoiced parts are then added to the original f_0 with 3 point median smoothing to reduce discontinuities in voicing boundaries. In addition, to alleviate edge artifacts, constant f_0 was added prior to and after the utterance. The pre-utterance f_0 value was set to the mean f_0 value calculated over the first half (in seconds) of the utterance; the post-utterance f_0 was set to the respective minimum. Finally, the interpolated $\log f_0$ contour is normalized to zero mean, unit variance as required by wavelet analysis. An example of an interpolated pitch contour is depicted in the second pane of Figure 1.

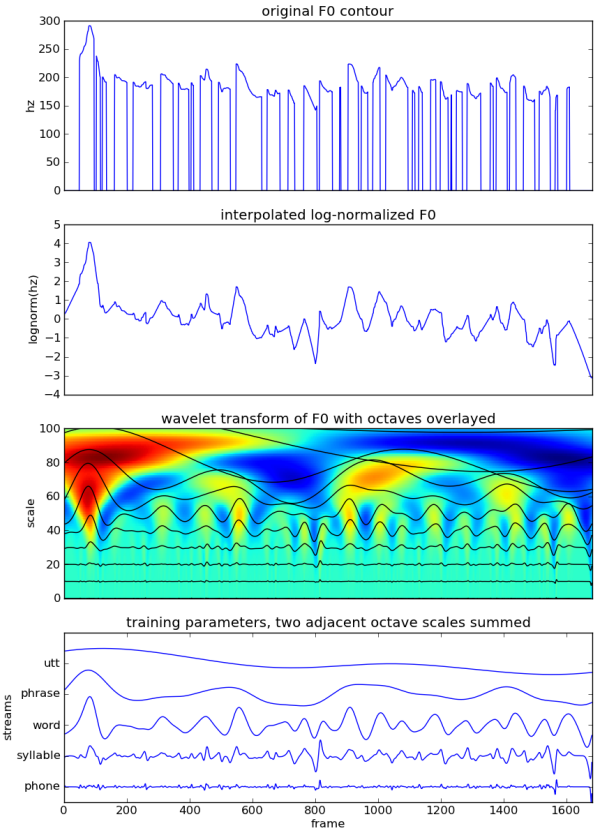

Figure 1: Example of f_0 parameterization. Top pane depicts the baseline method, *base*, in linear frequency scale; the second pane shows the interpolated baseline, *contf0*; third pane shows the continuous wavelet transform of the f_0 signal with the ten chosen scales separated by an octave (method *wave1*); the bottom pane shows the five scales that are merged from the continuous wavelet picture forming the basis of *wave5*

2.3. Wavelet based decomposition of f_0 contour

Wavelet transforms can be used to decompose a signal into frequency components similar to the Fourier transform. Although several alternatives exist, here we have chosen to use continuous wavelet transforms for f_0 decomposition. To define the wavelet transform, consider a (bounded) pitch contour f_0. The continuous wavelet transform $W(f_0)(\tau, t)$ of f_0 is defined by

$$W(f_0)(\tau, t) = \tau^{-1/2} \int_{-\infty}^{\infty} f_0(x) \psi \left(\frac{x - t}{\tau} \right) dx$$

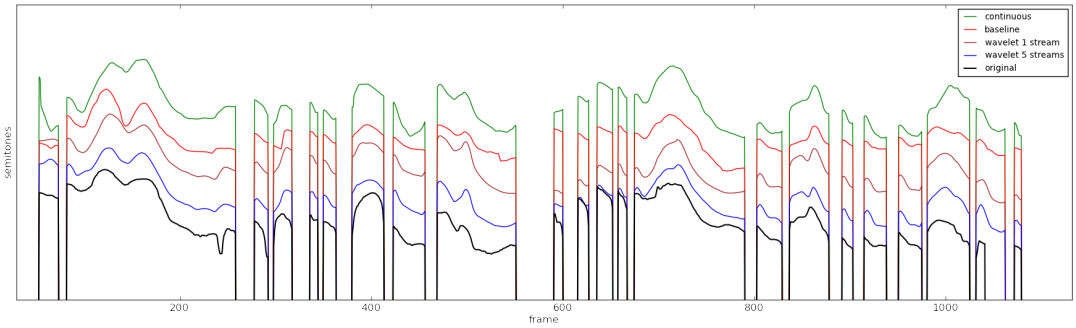

Figure 2: Example of synthesized f_0 contours with evaluated methods on a female corpus test utterance, overlaid three semitones apart.

where ψ is the Mexican hat mother wavelet. The original signal f_0 can be recovered from the wavelet representation $W(f_0)$ by inverse transform (for the proof, see [1, 5]):

$$f_0(t) = \int_{-\infty}^{\infty} \int_{0}^{\infty} W(f_0)(\tau, x)\tau^{-5/2}\psi\left(\frac{t-x}{\tau}\right) dx d\tau.$$

However, the reconstruction is incomplete, if all information on $W(f_0)$ is not availabe. Here, the decomposition and reconstruction is approximated by choosing ten scales, one octave apart. f_0 is represented by the wavelets as ten separate streams given by

$$W_i(f_0)(t) = W(f_0)(2^{i+1}\tau_0, t)(i+2.5)^{-5/2} \quad (1)$$

where $i = 1, \ldots, 10$ and $\tau_0 = 5$ ms, and the original signal is approximately recovered by

$$f_0(t) = \sum_{i=1}^{10} W_i(f_0)(t) \; + \; \epsilon(t) \quad (2)$$

where $\epsilon(t)$ is the reconstruction error. The reconstruction formula (2) is *ad hoc* and no attempts were made in this stage to optimize the computational efficiency. The accuracy of the reconstruction was evaluated by decomposing and reconstructing ten utterances spoken by a male and a female. The correlation between the original and the reconstructed f_0 signal was $99.7\,\%$ with root mean square reconstruction error of $1.03\ Hz$.

The continuous wavelet transform and ten distinct scales are shown in the third pane of the Figure 1. The scales 0 and 1 correspond to phone level (50 and 25 Hz), scales 2 and 3 correspond to syllable level (6 and 13 Hz), scales 4 and 5 show word level (1.6–3 Hz), scales 6 and 7 correspond to phrase level (0.4–0.8 Hz), and scales 8 and 9 correspond to utterance level. The adjacent scales are combined and shown in the bottom pane of the Figure 1. These five broad scales are separated by two octaves from each other. The correspondance of the prosodic levels of hierarchy and the wavelet scales is approximative and the wavelet scales are not adjusted to optimize the fit. Hence, e.g., not all the syllables have a duration that would fall in the "syllable scale".

3. Constructing the synthesis

3.1. Speech material

In order to carry out evaluation of the proposed f_0 modeling methods, two Finnish HMM-voices were trained, a male and a female one. The male database (MV) used is a traditional synthesis corpus, with rather carefully articulated set of 692 isolated sentences, while the female one (HK) is more diverse, consisting of 600 phonetically rich sentences as well as continuous prosodically rich read speech; 266 long sentences of fact and 607 sentences of diverse prose. 92 sentences of the male database was left out for evaluation purposes and 60 utterances of prose for the female. Both corpora have been tagged for word prominence on discrete scale ranging from 0 to 3, using acoustic features [13]. The prominence labels were used in both training and evaluation as contextual features. Thus the evaluation was not affected by TTS symbolic prosody prediction errors. In addition to word prominence, full context labels were generated with conventional features: quinphones with positional and length features of phones, syllables, word and phrases. Notably, more enriched labeling above word level would have been preferable for the current topic of modeling the prosodic hierarchy.

3.2. Parameterization of f_0 contours

Four different HMM-based statistical models for f_0 generation were compared. Synthesized f_0 contours based on these four and the original sentence f_0 are depicted in Figure 2.

3.2.1. base

A standard MSD model for f_0 is trained where each continuous f_0 passage between unvoiced segments is independently generated.

3.2.2. wave5

In the model *wave5*, five different f_0 components $w_1, \ldots, w_5$, defined by

$$w_i(t) = W_{2i-1}(f_0)(t) + W_{2i}(f_0)(t),$$

are independently trained by HMMs.

3.2.3. wave1

The different time scales correlate especially with their neighbors, so a plausible alternative would be to jointly model all the scales. This is done in *wave1* where one vector $V(t) = \{W_i(f_0)(t)\}_{i=1}^{10}$ contains the time scales.

3.2.4. *contf0*

Since the wavelet based methods *wave5* and *wave1* generate a continuous f_0 trajectory, and since interpolating the pauses in the training data improves the synthesized contours [14], an alternative, *contf0*, is offered where the unvoiced segments are interpolated in the same way as in the preprocessing of the wavelets.

3.3. HMM-training

The speech was parameterized with GlottHMM vocoder [9], yielding a 5-stream HMM structure: vocal tract spectrum LSFs and Gain (31 parameters), voice source spectrum LSFs (10), Harmonic-to-noise ratio (5) and $\log f_0$ (1). f_0 was then processed as described in the previous chapter. 5 streams (1 parameter each) for method wave5, 1 stream (10) for wave1 and one stream for continuous $\log f_0$. The baseline f_0 method was modeled as an MSD stream, others as continuous streams. With dynamic features further added, HMM training was perfomed in a standard fashion using HTS [16]. Stream weights affecting model alignment were set to zero for all streams except vocal tract spectrum LSFs and $\log f_0$. Decision tree clustering was perfomed individually for each stream without stream-dependent contextual question sets. Using the MDL criterion on decision tree building, the wave5 trees tended to become very large compared to baseline. Attempts were made to control the tree size with minimum leaf occupancy count, which was set to 10 on baseline MSD $\log f_0$ stream and 20, 25, 30, 60 and 70 for respective wave5 streams. In addition, MDL factor was set to 0.6 for $\log f_0$ stream and 1.5 for wave5 streams.

4. Evaluation

4.1. Evaluation data

The fundamental frequency parameters of the test utterances were generated from HMMs using original time alignments. For wavelet methods, the f_0 trajectories were constructed from generated scales using Equation (1). Voicing decision for continuous f_0 methods was based on the base MSD stream as well as mean and variance of f_0 for normalized wavelet methods.

The alignments were aquired by force-alignment method with the monophone models estimated during synthesis training. The synthesized sentences were checked manually for gross timing errors, and bad ones were excluded. The final MV test data consisted of 41 isolated utterances, spoken in the same formal style as the training data. By contrast, the HK test utterances consisted of 60 sentences of expressive prose.

4.2. Performance measures

The synthesized f_0 contours were compared to the original f_0 contours, estimated with GlottHMM, by measuring the correlation between the two curves and by calculating the root mean square error for each test utterance. Within an utterance, only the frames that were voiced with all methods were included. Also, due to frequent creaky voice with erratic pitch on original trajectories, the frames where the distance between original and at least one of the synthesized trajectories was more than 8 semitones, were excluded as outliers. It should be noted that these frames were completely excluded from the evaluation so that the comparisons were performed on exactly the same data sets. For the error calculation, the f_0 was converted to semitone scale with base 40 Hz. A Wilcoxon signed rank test was used to assess the statistical significance of the results.

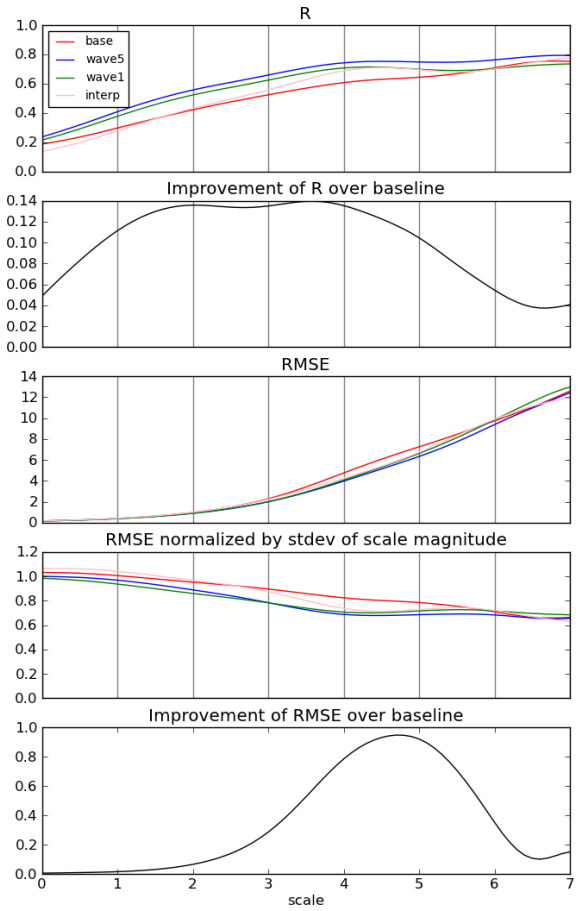

Figure 3: Evaluation results shown scale by scale. The top pane shows the correlations between the four synthesized contours and the original; second pane depicts the difference between the wavelet method *wave5* and *base*; third pane shows the absolute RMSE; in the fourth pane, the values are normalized by the variation at the scale; the bottom pane shows the difference in RMSE between the *wave5* and *base*.

4.3. Performance results

The correlations between the generated f_0 values and original contours showed significantly better performance for wavelet methods than for the baseline for both speakers. For the female data, the correlations over the test utterances were 0.76, 0.72, 0.72, and 0.68 for *wave5, wave1, contf0*, and *base*, respectively, as shown in Table 1. The *wave5* was better than *wave1* ($V = 1298$, $p < 0.05$), better than *contf0* ($V = 1324$, $p < 0.05$) and *base* ($V = 1445$, $p < 0.005$). In addition, the *wave1* was better than *base* ($V = 1329$, $p < 0.05$) but not significantly different

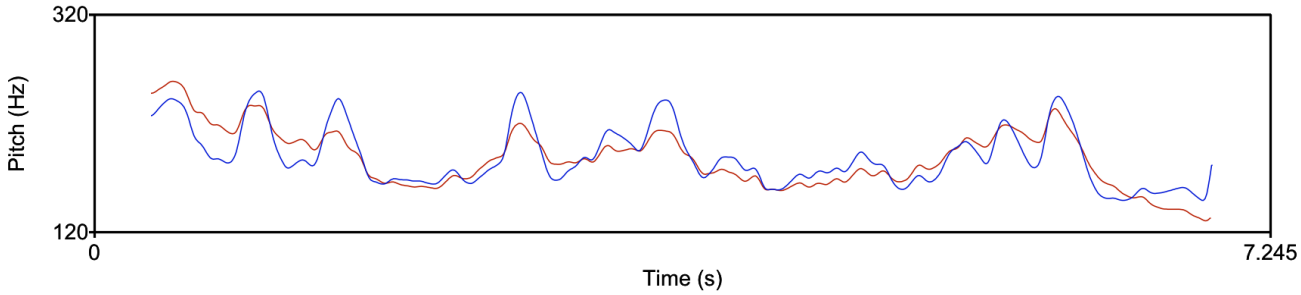

Figure 4: The reconstruction can be weighted to enhance the word level (blue curve) or the phrase level (red curve) intonation.

from *contf0* ($V = 1064$, $p > 0.1$). The *contf0* was marginally better than the *base* ($V = 702$, $p < 0.1$).

The male data showed similar patterns. The correlations over the test utterances were 0.85, 0.84, 0.81, and 0.81, respectively. The *wave5* was marginally better than *wave1* ($V = 288$, $p < 0.1$), better than *contf0* ($V = 129$, $p < 0.001$) and *base* ($V = 88$, $p < 0.001$). In addition, the *wave1* was better than *base* ($V = 136$, $p < 0.001$) and *contf0* ($V = 196$, $p < 0.005$). The *contf0* and the *base* were not significantly different ($V = 439$, $p > 0.1$).

Table 1: *A summary of the performance results of the syntheses. The means of the performance measures for each of the two data sets (female, male).*

	wave5	wave1	contf0	base
corr (F)	0.76	0.72	0.72	0.68
corr (M)	0.85	0.84	0.81	0.81
RMSE (F)	1.38	1.44	1.48	1.53
RMSE (M)	1.57	1.60	1.75	1.76

The root mean square error patterns are similar to the correlation results of the previous paragraphs. For the female data, the root mean square errors were 1.38, 1.44, 1.48, and 1.53 semitones for *wave5, wave1, contf0* and *base*, respectively. The *wave5* outperformed the *wave1* ($V = 1551$, $p < 0.001$), the *contf0* ($V = 1666$, $p < 0.001$, and the *base* ($V = 1781$, $p < 0.001$). The *wave1* and the *contf0* were statistically not different ($V = 1085$, $p > 0.1$), but the *wave1* was better than the *base* ($V = 1419$, $p < 0.005$). The *contf0* was better than *base* ($V = 599$, $p < 0.01$). For the male data, the root mean square error was 1.57, 1.60, 1.75, and 1.76 semitones for *wave5, wave1, contf0* and *base*, respectively. The *wave5* was not different from the *wave1* ($V = 307$, $p > 0.1$) but was better than the *contf0* ($V = 143$, $p < 0.001$) and the *base* ($V = 96$, $p < 0.001$). The *wave1* outperformed both the *contf0* ($V = 206$, $p < 0.005$) and the *base* ($V = 145$, $p < 0.001$). Finally, the *contf0* and *base* did not differ significantly ($V = 433$, $p > 0.1$).

4.4. Temporal scale analysis of the results

In Figure 3, the performance measures over the female test sentences are decomposed to the scale-wise components. Overall, the *wave5* is better than the baselines at all scales. However, the difference is pronounced for the middle scales.

5. Discussion and conclusions

The results of the objective evaluation are in line with previous research. Continuous f_0 modeling is found significantly better than the standard HTS method. On male voice, the synthesis of f_0 is very accurate, suggesting that existing methods are capable of modeling higher level structures to an adequate degree, given consistent style and accurate labels of word prominence. Consequently, the differences between evaluated methods are rather small, though the wavelet based methods provide some gains. As expected, the performance of all evaluated methods is lower on female voice due to difficult test utterances of continuous expressive prose, and also possibly due to more errors in f_0 estimation during analysis. Here, the individually modeled wavelet scales provide a large improvement. However, subjective evaluation is still required for final conclusions.

Overall, the results suggest that the proposed method largely solves the fragmentation problem caused by simultaneous decision tree clustering of all levels of prosodic hierarchy. Yet, somewhat contrary to expectations the improvements seem larger on word level and syllable level than on phrase level. Although technical problems of higher scales affected by boundary effects on wavelet analysis may have an effect, this mainly highlights the need for new contextual features on supraword level, beyond position and number. With the proposed method the features representing for instance constituent structure, phrase type and utterance modality could actually have an effect on the synthesized prosody.

The wavelet decomposition offers a possibility of adjusting the weights of individual scales prior to reconstruction. This could have potential applications in speaking style modification. For example, informal listening suggested that increasing the weight of the word level makes the synthesized speech sound more resolute and perhaps more intelligible, while listening longer passages is less displeasing when phrase level is emphasized. Moreover, moderate modifications do not seem to have adverse effect on naturalness. Figure 4 presents an example of this type of modification. Local weighting within utterance could also be applied for e.g. emphasis reproduction. Rapid adaptation of speaking style based on transform of the scale weights alone could also be considered.

The current paper has presented a novel method of f_0 modeling based on wavelet decomposition. Many open questions remain. Selection of scales and model structure were made based on intuition alone, no other wavelets beyond mexican hat were considered, neither more popular discrete wavelet transform.

Also, while the proposed method seems quite suitable for the current HMM-synthesis framework, it is deeply unsatisfying to model utterance level f_0 contour with inherently sub-segmental models, when the discrete cosine transform or discrete wavelet transform could represent the level with only a few coefficients.

6. Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement n^{o} 287678 and the Academy of Finland grants 128204 and 125940.

7. References

[1] Daubechies, I., "Ten lectures on wavelets", Philadelphia, SIAM, 1992.

[2] Fujisaki, H., Hirose, K., Halle, P., and Lei, H., "A generative model for the prosody of connected speech in Japanese", Ann. Rep. Eng. Research Institute 30: 75–80, 1971.

[3] Kruschke, H. and Lenz, M., "Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis", in Proc. Eurospeech'03, 4, pp. 2881-2884, Geneva, 2003.

[4] Lei, M., Wu, Y. J., Ling, Z. H., and Dai, L. R., "Investigation of prosodic F_0 layers in hierarchical F_0 modeling for HMM-based speech synthesis", Proc. IEEE Int. Conf. Signal Processing (ICSP) 2010, 613–616.

[5] Mallat, S., "A wavelet tour of signal processing", Academic Press, San Diego, 1998.

[6] Mishra, T., van Santen, J., and Klabbers, E., "Decomposition of pitch curves in the general superpositional intonation model", Speech Prosody, Dresden, Germany, 2006.

[7] Öhman, S., "Word and sentence intonation: a quantitative model", STLQ progress status report, 2–3:20–54, 1967.

[8] L. Rabiner, "On the use of autocorrelation analysis for pitch detection", IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 25, no. 1, pp. 24–33, 1977.

[9] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", IEEE Trans. on Audio, Speech, and Lang. Proc., 19(1):153–165, 2011.

[10] van Santen, J. P. H., Mishra, T., and Klabbers, E., "Estimating phrase curves in the general superpositional intonation model", Proc. 5th ISCA speech synthesis workshop, Pittsburgh, 2004.

[11] Stan, A. and Giurgiu, M., "A Superpositional Model Applied to F0 Parameterization using DCT for Text-to-Speech Synthesis", Proceedings of 6th Conference on Speech Technologoy and Human-Computer Dialogue (SpeD), 2011, 1–6, Brasov.

[12] Teutenberg, J., Watson, C. I., and Riddle, P., "Modelling and synthesising F0 contour with the discrete cosine transform", ICASSP 2008, 3973–3976, 2008.

[13] Vainio, M., Suni, A., and Sirjola, P., "Accent and prominence in Finnish speech synthesis", Proc. 10th Int. Conf. Speech and Computer (Specom 2005), 309–312.

[14] Yu, K. and Young, S., "Continuous F0 Modeling for HMM based statistical parpametric speech synthesis", Trans. Audio, Speech and Lang. Proc., 19:5, 1071–1079, 2011.

[15] Zen, H. and Braunschweiler, N., "Context-dependent additive log F0 model for HMM-based speech synthesis", Proc. Interspeech 2009: 2091–2094.

[16] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., Tokuda, K., 2007. The HMM-based speech synthesis system (HTS) version 2.0. In: SSW6. pp. 294–299.