Continuous wavelet transform for analysis of speech prosody

Martti Vainio, Antti Suni, and Daniel Aalto

Institute of Behavioural Sciences (SigMe Group), University of Helsinki, Finland

martti.vainio@helsinki.fi, antti.suni@helsinki.fi, daniel.aalto@helsinki.fi

Abstract

Wavelet based time frequency representations of various signals are shown to reliably represent perceptually relevant patterns at various spatial and temporal scales in a noise robust way. Here we present a wavelet based visualization and analysis tool for prosodic patterns, in particular intonation. The suitability of the method is assessed by comparing its predictions for word prominences against manual labels in a corpus of 900 sentences. In addition, the method's potential for visualization is demonstrated by a few example sentences which are compared to more traditional visualization methods. Finally, some further applications are suggested and the limitations of the method are discussed.

Index Terms: continuous wavelet transform; speech prosody; intonation analysis; prominence

1. Introduction

The assumption that prosody is hierarchical is shared by phonologists and phoneticians alike. There are several accounts for hierarchical structure with respect to speech melody: In the tone sequence models which interpret the f_0 contour as a sequence of tonal landmarks of peaks and valleys (e.g. [15]) the hierarchy is mainly revealed at the edges or boundaries of units whereas in superpositional accounts (e.g., [13, 6]) it is seen as a superposition of different levels at each point of the contour. The problem with the tone sequence models stems from their phonological nature which requires a somewhat discretized view of the continuous phonetic phenomena. The superpositional accounts suffer, conversely, from the lack of signal based categories that would constrain the analysis in a meaningful way. Both models suffer from being disjointed from perception and require *a priori* assumptions about the utterances.

Wavelets emerged independently in physics, mathematics, and engineering, and are currently a widely used modern tool for analysis of complex signals including electrophysiological, visual, and acoustic signals [5]. In particular, the wavelets have found applications in several speech prosody related areas: The first steps of the signal processing by the auditory periphery are well described by models that rely on wavelets [23, 22, 17]; they are used in a robust speech enhancement in noisy signals with unknown or varying signal to noise ratio, in automatic speech segmentation, and in segregation along various dimensions of speech signal in a similar way as mel-cepstral coefficients [2, 1, 8, 9]; the multiscale structure of the wavelet transform has been taken advantage of in musical beat tracking [19]. The quantitative analysis of speech patterns through wavelets might also be relevant for understanding the cortical processing of speech (e.g. [3, 14, 7]).

In the present paper, we apply the wavelet methods to recorded speech signals in order to extract prosodically important information automatically. Here, only the fundamental frequency of the speech signal is analyzed by wavelets although similar analysis could be performed to any prosodically relevant parameter contour (e.g., the intensity envelope contour or a speech rate contour) or even the raw speech signal itself.

The analysis of intonation by wavelets is not a new idea. Discrete wavelet analysis with Daubechies mother wavelets was the key component in automatically detecting the correct phrasal components of synthesized f_0 contours of the Fujisaki model further developed under the name general superpositional model for intonation proposed by van Santen et al. [21, 12]. Continuous wavelet transforms with Mexican hat mother wavelet have been used for Fujisaki accent command detection by Kruschke and Lenz [10]. Overall, previous work with wavelets and f_0 have been mainly concerned with utilizing wavelets as a part of model development or signal processing algorithm, instead of using the wavelet presentation itself.

In Finnish, the prosodic word is an important hierarchical level and the prominence at that level reveals much of the syntactically and semantically determined relations within the utterances. We have successfully used a four level word prominence in text-to-speech synthesis in both Finnish and English [20] and the automatic detection of word prominence is a prerequisite for building high quality speech synthesis. In relation to both a tone sequence and superpositional accounts the sucesfull detection of word prominence would be related to distinguishing the accentedness of the unit as well as the magnitude of the accent.

Using an inherently hierarchical analysis we can do away with a fixed model and try to directly link acoustical features of an utterance to the perceived prominences within the utterance. In order to evaluate the wavelet analysis we calculated CTW based prominences for about 7600 separate words in 900 utterances previously annotated by human labelers and compared various wavelet and f_0 based features with each other. In this paper we first discuss the CWT and its application to f_0 and then show the quantitative evaluation followed by discussion and conclusion.

2. Continuous wavelet transform

The continuous wavelet transform (CWT) can be constructed for any one-dimensional or multidimensional signal of finite energy. In addition to the dimensions of the original signal, CWT has an additional dimension, scale, which describes the internal structure of the signal. This additional dimension is obtained by convolving the signal by a mother wavelet which is dilated to cover different frequency regions [5]. The CWT is similar to the windowed Fourier transform: the CWT describes the time-

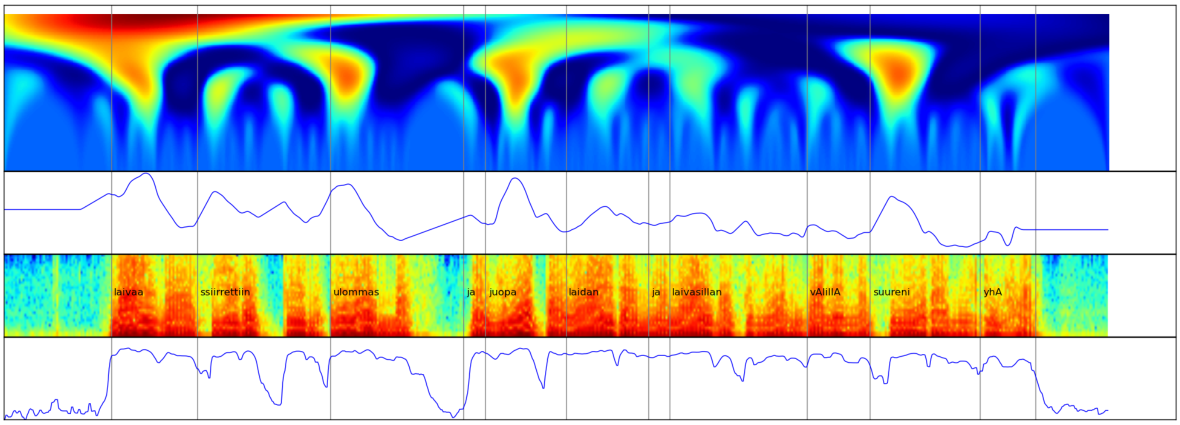

Figure 1: Different analyses aligned temporally. Top pane depicts the continuous wavelet transform with Mexican hat mother wavelet of f_0, second pane shows the interpolated f_0 contour; third pane shows spectrogram of the speech signal; the bottom pane shows gain. The light gray vertical lines show the word boundaries. The text superposed to the third pane transcribes the uttered words (The ship was moved outwards and the gap between the board of the ship and the gangplank got wider, still.)

frequency behaviour of the signal and the signal can be reconstructed from the CWT by inverse wavelet transform. We use here a Mexican hat shaped mother wavelet which corresponds formally to the second derivative of the Gaussian, see pages 76–78 in [11]. In the Figure 1, the top pane shows the CWT of the f_0 contour shown in the second pane. The peaks in f_0 curve show up in the CWT as well, but the size of the peaks in the wavelet picture depends on the local context: the higher at the picture, or in other words, the coarser the scale, the slower the temporal variations and the larger the temporal integration window. Although several hierarchical levels emerge, the quantitative evaluation of the suitability of the CWT to prosodic analysis is only performed on word level. Note that in Finnish, content words have a fixed stress on the first syllable, clearly visible in the Figure 1. The third and fourth panes show the spectrogram and the intensity envelope of the same utterance. The time scales in the wavelet picture range from the 67 Hz as finest to less than 1 Hz as coarsest.

3. Quantitative evaluation

A visualization tool cannot be evaluated quantitatively as a whole. However, if the different temporal scales reflect perceptually relevant levels of prosodic hierarchy, the representation of f_0 at any scale should correlate with judgements of the relative prominence at that particular level. This hypothesis is tested at the level of prosodic word. Although word prominence is signaled by f_0, it is, to large extent, signaled by other means as well including intensity, duration, word order, and morphological marking. Hence, the f_0 based prominence annotation is compared to a simple baseline f_0 prominence annotator and to the labels obtained from phonetically trained listeners.

3.1. Recorded speech data

The evaluation data consisted of 900 read sentences by a phonetically trained, native female speaker of Finnish. Linguistically, the sentences represented three different styles: modern standard scientific Finnish, standard Finnish prose, and phonetically rich sentences covering the Finnish phonemes. The sentences were recorded using high quality condenser microphone in a sound proof studio, digitized, and stored on a computer hard drive. The mean durations of the sentences had average durations of 6.1 s, 3.5 s, and 3.8 s. The total duration amounted to 1h 1 min. Acoustic features were extracted of the utterances with GlottHMM [16], and then the utterances were aligned with the text.

3.2. Fundamental frequency extraction

The fundamental frequency of the test utterances were extracted by GlottHMM speech analysis and synthesis software. In GlottHMM analysis, the signal is first separated to vocal tract and glottal source components using inverse filtering, and the f_0 is then extracted from the differentiated glottal signal using autocorrelation method. Parameters concerning voicing threshold and admissible range of f_0 values were tuned manually for the current speaker. While GlottHMM performs some post-processing on analyzed f_0 trajectories, deviations from perceived pitch remain, particularly in passages containing creaky voice. Thus, f_0 values were first transformed to logarithm scale and then all values lower than 2 standard deviations below the mean of log f_0 were removed.

The unvoiced segments of the speech and the silent intervals make the direct wavelet analysis impossible since f_0 is not well defined for these segments. Hence, the unvoiced gaps were filled using linear interpolation. Additionally, to alleviate edge artifacts, the continuous f_0 contour was extended over the silent beginning and end intervals by replacing the former by the mean f_0 value (logarithmically scaled) over the first half of the completed f_0 contour, and the latter by the mean over the second half. Then the f_0 curve was filtered by a moving average Hamming window of length 25 ms and finally normalized to zero mean and unity variance.

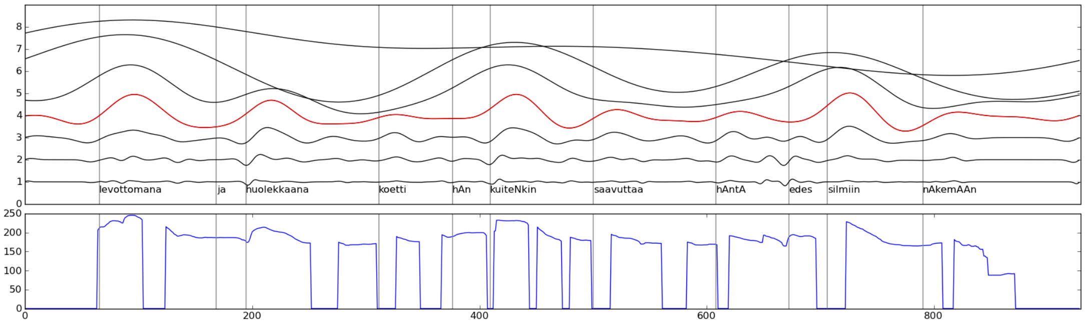

Figure 2: The word prosody scale is chosen from a discrete set of scales with ratio 2 between ascending scales as the one with the number of local maxima as close to the number of words in the corpus as possible. The upper pane shows the representations of f_0 at different scales. The word level (4.2 Hz; see text) is drawn in red. The lower pane shows the f_0 curve. The abscissa shows the frame count from the beginning of the utterance (5 ms frame duration).

3.3. Baseline annotation based on f_0 signal

For each word in the evaluation data, we extracted two common measurements from the preprocessed and normalized f_0 signal, the maximum value observed during word (BMax) and the maximum minus minimum (BRange). The measurements were not further processed, despite the scale differences compared to manual annotation, as only correlation was being tested.

3.4. CWT annotation based on f_0 signal

The CWT transform was first perfomed with one scale per octave, with finest scale being 3 frames or 15 ms. Then, the scale of interest for word prominence was selected as the one with positive peak count closest to the number of words (see Figure 2; the word scale corresponds to 4.2 Hz in the current data). This is intuitively suitable for Finnish, with relatively few unaccented function words. Three wavelet based measurements were then extracted for each word, height of the first local maximum (WPeak) as well as the same two measurements as in f_0 baseline (WMax, WRange). If the word contained no maxima, then the prominence of the word was set to zero. Note that the peak method is not applicable to raw F0, as the noisier contour contains many peaks. More complex measurements were experimented with, such as averaging over multiple scales, but with only moderate success.

3.5. Prominence labeling

Ten phonetically trained listeners participated in prominence labeling. The listeners were instructed to judge the prominence of each word in a categorical scale: 0 (unaccented, reduced); 1 (perceivably accented but no emphasis); 2 (accented with emphasis); 3 (contrastive accent). The listeners reported to have based their judgements mainly on listening and secondarily to the available Praat analyses of pitch, intensity, and spectrogram. Every listener labeled 270 sentences in such a way that every sentence was labeled by three listeners. The prominence of a word was set to the average of the three judgements.

3.6. Statistical analysis

The two baseline annotations and the three wavelet based annotations were compared to the listeners' judgements of word prominence by linear regression analysis. The amount of variance explained (R squared) by the regression model was used as an indicator for the goodness of the used measure.

3.7. Results

The baseline measure $BMax$ has a strong correlation to the prominence judgements with 37 % of the variance explained. The other baseline measure $BRange$ explained 36 % of the variance. The wavelet based measures fitted better to the data: $WMax$ and $WRange$ explained 47 % and 39 % of the variance, respectively. The more involved measures $WPeak$ explained 53 % of the variance.

4. Discussion

The results of the evaluation show that it is fairly straightforward to extract prosodically relevant information form the CWT analysis. In this case it was at the level of prosodic word (which in Finnish correponds well with the grammatical word). As can be seen in Figures 1 and 2, there are other levels both above and below the word that are relevant and if discretized, form a hierarchical tree which can be further exploited for instance in text-to-speech synthesis. However, such an analysis is not free of problems. For instance, the temporal scale corresponding to syllables becomes coarser (higher levels in the Figure 1) when the speech slows down, as is the case in e.g. pre-pausally.

What is important to notice here is that the CWT analysis – as applied to the pitch contour – takes into account both the f_0 level and its temporal properties as cues for prominence. Although we only used one level it is the analysis as a whole that we are interested in. As mentioned earlier, the wavelet analysis can be done on any prosodically relevant signal either alone or jointly – although multidimensional may no longer be easily visualizable.

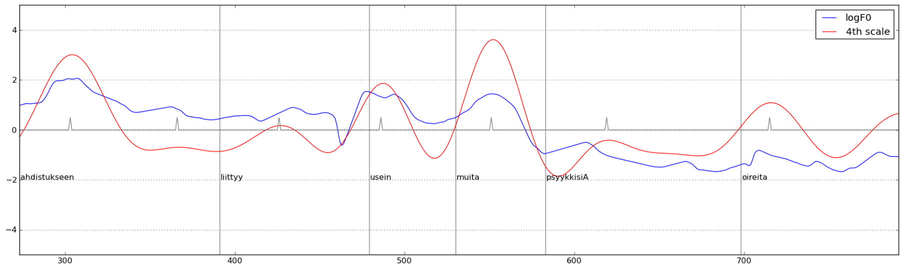

Figure 3: Comparison of selected word scale and original f_0 contour with detected peaks marked with gray triangles. Observe that the wavelet contour is free of noise and declination trend.

5. Conclusion

Continuous wavelet transfrom, a standard mathematical tool for simultaneous analysis and visualization of various temporal scales of a signal, is applied to f_0 signal of recorded speech. At the temporal scale corresponding to prosodic word, the local maxima correlate strongly with the listeners' judgements on the perceived word prominence. This is taken as evidence that the small and large scale contributions induced by segmental micro-prosody and phrasal intonation components are effectively removed by the analysis. Moreover, a hierarchical structure emerges which is easily visible and has similarities with the classical description of prosodic structure through a prosodic tree. Unlike other hierarchical models of prosody, the structure rises directly from the signal with no assumptions on the f_0 model.

Some interesting future directions could include building a 'spectrogram of prosody' -visualization tool combining spectrogram and prosody in the same picture, attempting to discretize the hierarchical structure for higher level applications, applying the decomposed prosodic features for TTS prosody models, studying other prosodic features such as energy by CWT, and, finally, exploring the relationship between the CWT analyses and human auditory processing.

6. Acknowledgments

The research leadgin 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 (projects 135003 LASTU programme, 1128204, 128204, 125940). We would also like to thank Heini Kallio for collecting the prominence data.

7. References

[1] Alani, A. and Deriche, M., "A novel approach to speech segmentation using the wavelet transform", 5th Int. Symposium on Signal Processing and its Applications, Brisbane, 1999.

[2] Bahoura, M., "Wavelet speech enhancement based on the Teager energy operator", IEEE Signal Processing Letters, 8(1):10–12, 2001.

[3] Bradley, A. P. and Wilson, W. J., "On wavelet analysis of auditory evoked potentials", Clinical neurophysiology, 115:1114-1128, 2004.

[4] Chi, T., Ru, P., and Shamma, S. A., "Multiresolution spectrotemporal analysis of complex sounds", J. Acoust. Soc. Am. 118(2):887–906, 2005.

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

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

[7] Giraud, A. and Poeppel, D., "Cortical oscillations and speech processing: emerging computational principles and operations", Nature Neuroscience, 15:511–517, 2012.

[8] Hu, G. and Wang, D., "Segregation of unvoiced speech from nonspeech interference", J. Acoust. Soc. Am., 124(2): 1306–1319, 2008.

[9] Irino, T. and Patterson, R. D., "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: the stabilised wavelet-Mellin transform", Speech Communication, 36:181–203, 2002.

[10] 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.

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

[12] Mishra, T., van Santen, J., and Klabbers, E., "Decomposition of pitch curves in the general superpositional intonation model",

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

[14] Petkov, C. I., O'Connor, K. N., and Sutter, M., L., "Encoding of illusory continuity in primary auditory cortex", Neuron, 54: 153–165, 2007.

[15] Pierrehumbert, J., "The phonology and phonetics of English intonation", PhD Thesis, MIT, 1980.

[16] 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.

[17] Reimann, H. M., "Signal processing in the cochlea: the structure equations", J. Mathematical Neuroscience, 1(5):1–50, 2011.

[19] Smith, L. M. and Honing, H., "Time-Frequency representation of musical rhythm by continuous wavelets", J. Mathematics and Music, 2(2):81–97, 2008.

[20] Suni, A., Raitio, T., Vainio, M., and Alku, P., "The GlottHMM entry for Blizzard Challenge 2012 – hybrid approach", in Blizzard Challenge 2012 Workshop, Portland, Oregon, 2012.

[21] 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.

[22] Yang, X., Wang, K., and Shamma, S., "Auditory representation of acoustic signals", IEEE Trans. Information theory, 38:824–839, 1992.

[23] Zweig, G., "Basilar membrane motion", Cold Spring Harbor Symposia on Quantitative Biology, 40:619–633, 1976.