Hierarchical Representation and Estimation of Prosody using Continuous Wavelet Transform

Antti Suni[a,b], Juraj Šimko[a], Daniel Aalto[c,d,a], Martti Vainio[a,*]

[a] *Institute of Behavioural Sciences, University of Helsinki*
[b] *Department of Signal Processing and Acoustics, Aalto University*
[c] *Communication Sciences and Disorders, Faculty of Rehabilitation Sciences, University of Alberta*
[d] *Institute for Reconstructive Sciences in Medicine (iRSM), Misericordia Hospital, Edmonton*

Abstract

Prominences and boundaries are the essential constituents of prosodic structure in speech. They provide for means to chunk the speech stream into linguistically relevant units by providing them with relative saliences and demarcating them within utterance structures. Prominences and boundaries have both been widely used in both basic research on prosody as well as in text-to-speech synthesis. However, there are no representation schemes that would provide for both estimating and modelling them in a unified fashion. Here we present an unsupervised unified account for estimating and representing prosodic prominences and boundaries using a scale-space analysis based on continuous wavelet transform. The methods are evaluated and compared to earlier work using the Boston University Radio News corpus. The results show that the proposed method is comparable with the best published supervised annotation methods.

Keywords: phonetics, prosody, speech synthesis, wavelets

1. Introduction

Two of the most primary features of speech prosody have to do with chunking speech into linguistically relevant units above the segment and the relative salience of the given units; that is, boundaries and prominences, respectively. These two aspects are present in every utterance and are central to any representation of speech prosody. Arrangement of prominence patterns and placement of boundaries reflect the hierarchical structure of speech, i.e., gradual nesting of units, segments within syllables, syllables within (prosodic) words, words within

[*] Corresponding author

Email addresses: antti.suni@helsinki.fi (Antti Suni), juraj.simko@helsinki.fi (Juraj Šimko), aalto@ualberta.ca (Daniel Aalto), martti.vainio@helsinki.fi (Martti Vainio)

phrases, phrases within utterances and beyond [1]. Borders between adjoining units of higher order – words, phrases – present affordances for prosodic breaks of different types and strengths. Attention of the listener can be selectively drawn to individual units within the hierarchy; prominent syllables mark lexical stress, prominent words signal focus, etc.

In speech, boundaries are usually signalled by a local reduction in one or more signal characteristics (such as intensity or pitch) at a border spanning several hierarchical levels. In a complementary fashion, prominence is typically associated with an increase in some or all of these signal properties, typically associated with a particular hierarchical level.

This simple insight suggests that these prosodic constituents could be represented within a uniform methodology that identifies both prominence and boundaries as complementary phenomena manifested in speech signals. Such a methodology would be beneficial to both basic speech research and speech technology, especially speech synthesis and recognition. At the same time, to be useful for data oriented research and technology, the annotation system should strive towards being unsupervised as opposed to the systems that rely on humans, either directly labelling speech data [2] or providing a manually labeled training set used for training the system.

Ideally, the system should approach human-like performance but without the variability of human labellers caused by complex interactions between the top-down and bottom-up influences. In order to achieve that we propose here a system based on Continuous Wavelet Transform (CWT) that (1) approximates human processing of a complex signal relevant for identifying prominence and boundaries, and (2) is capable of representing the speech signal in a manner that captures the hierarchical nature of prosodic signalling.

In this paper we present a hierarchical, time-frequency scale-space analysis of prosodic signals (e.g., fundamental frequency, energy, duration) based on the CWT. The presented algorithms can be used to analyse and annotate speech signals in an entirely unsupervised fashion. The work stems from the need to annotate speech corpora automatically for text-to-speech synthesis (TTS) [3] and the subject matter is partly examined from that point of view. However, the presented representations should be of interest to anyone working on speech prosody.

Wavelets extend the classical Fourier theory by replacing a fixed window with a family of scaled windows resulting in scalograms, resembling the spectrogram commonly used for analysing speech signals. The most interesting aspect of wavelet analysis with respect to speech is that it resembles the perceptual hierarchical structures related to prosody. In scalograms, speech sounds, syllables, (phonological) words, and phrases can be localised precisely in both time and frequency (scale). This would be considerably more difficult to achieve with traditional spectrograms. Furthermore, the wavelets give natural means to discretise and operationalise the continuous prosodic signals.

Figure 1 shows how the hierarchical nature of speech can be captured in a time-frequency scale-space by CWT of a composite prosodic signal of an English utterance. The scalogram is shown as a heat map in the top part of the figure

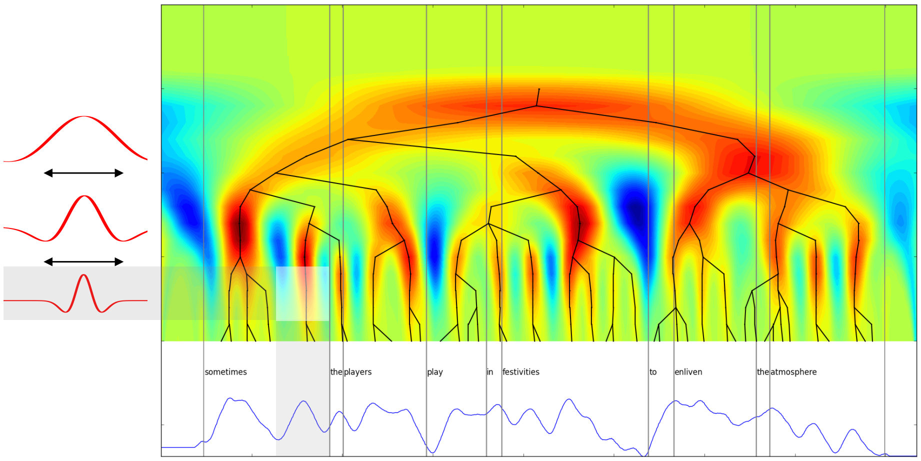

Figure 1: An illustration of a CWT based analysis of a composite prosodic signal combining energy, f_0 and word duration (bottom) of an English utterance ‘Sometimes the players play in festivities to enliven the atmosphere”. The hierarchical tree structure is highlighted in black. Mother wavelets corresponding to syllables, prosodic words, and phrases are depicted on the left. See text for more detail.

above the signal contour in blue. The scalogram is constructed from multiple scale functions (see also Fig. 2). Each of these scale functions is a convolution of the original signal and a dilated, i.e., scaled version of the mother wavelet (the Mexican hat wavelet in this case). Three examples of the scaled wavelets are shown to the left of the scalogram; as we can see, scalogram results from the convolution with progressively more and more scaled up – wider and higher – wavelets. The convolution operator depicts “similarities” between the two convoluted functions, signal and the wavelet. As the highlighted area in the figure illustrates, a local similarity in shape of the signal with the dilated wavelet leads to a higher value of the scale function (red area in the heat map); the most dissimilar portions of the signal (valleys compared to peak-like shape of the wavelets) yield negative values of the scale function shown in blue.

The tree structure superimposed in black over the scalogram in Fig. 1 joins the red areas of high similarity with differently scaled wavelets, i.e., depicts the hierarchy of portions of the signal that are “prominent” at various scales. The hierarchical utterance structure has served as a basis for modelling the prosody, e.g., speech melody, timing, lexical stress, and prominence structure of the synthetic speech.

Controlling prosody in synthesis has been based on a number of different theoretical approaches stemming from both phonological considerations as well as phonetic ones. The phonologically based ones stem from the so called Autosegmental Metrical theory [4] which is based on the three-dimensional phonology developed in [5, 6] as noted in [7]. These models are sequential in nature though a hierarchical structure is explicitly referred to for certain features of the models (e.g., break indices in ToBI [2]). The more phonetically oriented hierarchical models are based on the assumption that prosody – especially intonation – is truly hierarchical in a super-positional and parallel fashion.

Actual models capturing the superpositional nature of intonation were first proposed in [8] by Öhman, whose model was further developed by Fujisaki *et al.*[9, 10] as a so called command-response model which assumes two separate types of articulatory commands; accents associated with stressed syllables superposed on phrases with their own commands. The accent commands produce faster changes which are superposed on slowly varying phrase contours. Several superpositional models with a varying degree of levels have been proposed since Fujisaki [11, 12, 13, 14]. Superpositional models attempt to capture both the chunking of speech into phrases as well the highlighting of words within an utterance. Typically smaller scale changes, caused by e.g., the modulation of the airflow (and consequently the f_0) by the closing of the vocal tract during certain consonants, are not modelled.

Prominence is a functional phonological phenomenon that signals syntagmatic relations of units within an utterance by highlighting some parts of the speech signal while attenuating others. Thus, for instance, some of the syllables within a word stand out as stressed [15]. At the level of words prominence relations can signal how important the speaker considers each word in relation to others in the same utterance. These often information based relations range from simple phrasal structures (e.g., prime minister, yellow car) to relating utterances to each other in discourse as in the case of contrastive focus (e.g., "Where did you leave your car? No, we WALKED here."). Although prominence impressions might be continuous, they may serve categorical functions. Thus, the prominence can be categorised [16, 17] in, e.g, four levels where the first level stands for words that are not stressed in any fashion prosodically to moderately stressed and stressed and finally words that are emphasised (as the word WALKED in the example above). These four categories are fairly easily and consistently labeled even by non-expert listeners [18]. In sum, prominence functions to structure utterances in a hierarchical fashion that directs the listener's attention in a way which enables the understanding of the message in an optimal manner. However, prominent units – be they words or syllables – do not by themselves demarcate the speech signal but are accompanied by boundaries that chunk the prominent and non-prominent units into larger ones: syllables to (phonological) words, words to phrases, and so forth. Prominence and boundary estimation have been treated as separate problems stemming from different sources in the speech signals.

As functional – rather than formal, purely signal-based – prosodic phenomena, prominences and boundaries lend themselves optimally to statistical modelling (traditionally by supervised methods). The actual signalling of prosody in terms of speech parameters is extremely complex and context sensitive: the form follows function in a complex fashion. Capturing prominence and boundaries in terms of one-dimensional values reduces representational complexity of speech annotations in an advantageous way. In a synthesis system this reduction occurs at a juncture that is relevant in terms of both representations and data scarcity.

The complex feature set that is known to effect the prosody of speech can be narrowed to a few categories or a single continuum from dozens of context sensitive features, such as e.g, part-of-speech and whatever can be computed from the input text. Taken this way, both prominence and boundaries can be viewed as abstract phonological functions that impact the phonetic realisation of the speech signal predictably despite a possible considerable phonetic variation.

The perceived prominence of a given word in an utterance is a product of many separate sources of information; mostly signal-based although other linguistic, top-down factors have been shown to modulate the perception [19, 18, 20, 15, 16, 21]. Typically a prominent word is accompanied with a clearly audible f_0 movement, the stressed syllable is longer in duration, and its intensity is higher. However, because of the combination of the bottom-up and top-down influences, and their manifestation in the hierarchical character of speech discussed above, estimating prominences automatically is not straight-forward and a multitude of different estimation algorithms have been suggested (see Section 3 for more detail).

In what follows we present recently developed methods for automatic prominence estimation and boundary detection based on CWT (Section 2) which allow for fully automatic and unsupervised means to estimate both (word) prominences and boundary values from a hierarchical representation of speech (see [22, 23, 24] for earlier work). The main insight in this methodology is that both prominences and boundaries can be treated as arising from the same sources in the (prosodic) speech signals and estimated with exactly the same methods. These methods, then, provide for a uniform representation for prosody that is useful in both speech synthesis and basic phonetic research. These representations are purely computational and thus objective. It is – however – interesting to see how the proposed hierarchical method relates to annotations provided by humans as well as earlier attempts at the problem (Section 3).

2. Methods

Wavelets are used in a great variety of applications for effectively compressing and denoising signals, to represent the hierarchical properties of multidimensional signals like polychromatic visual patterns in image retrieval, and to model optical signal processing of visual neural fields [25, 26]. In speech and auditory research there is also a long history going back to the 1970's [27, 28, 29, 30, 31, 32, 33, 34, 35]. A recent summary of wavelets in speech technology can be found in [36].

In speech synthesis context, wavelets have been used mainly for parameter estimation [37, 38, 39] but never as a full modelling paradigm. In the HMM based synthesis framework, decomposition of f_0 to its explicit hierarchical components during acoustic modelling has been investigated in [40, 41]. These approaches rely on exposing the training data to a level-dependent subset of questions for separating the layers of the prosodic hierarchy. The layers can then be modelled separately as individual streams [40], or jointly with adaptive training methods [41].

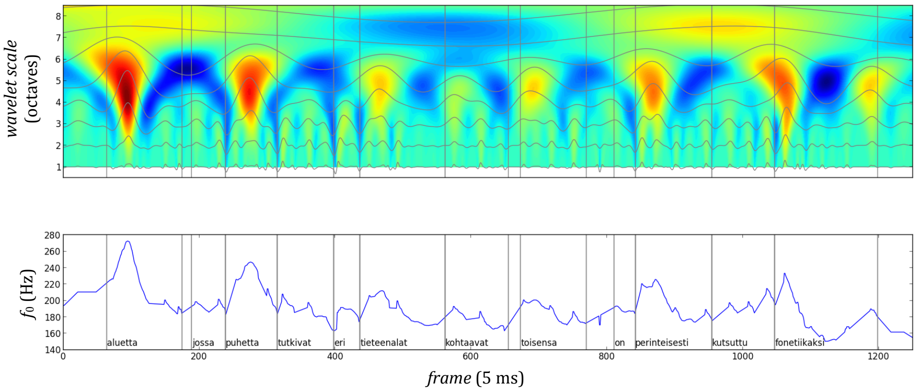

Figure 2: CWT of the f_0 contour of a Finnish utterance. The lower pane shows the (interpolated) contour itself as well as orthographic words (word boundaries are shown as vertical lines in both panes). The upper pane shows the wavelet transform as well as eight separated scales (grey lines) ranging from segmentally influenced perturbation or microprosody (lowest scale) to utterance level phrase structure (the highest level).

Figure 2 shows a CWT of interpolated f_0 contour of the Finnish utterance "Aluetta, jossa puhetta tutkivat eri tieteenalat kohtaavat toisensa on perinteisesti kutsuttu fonetiikaksi", (The area where the sciences interested in speech meet each other has been traditionally called phonetics.). The lower pane shows the contour itself as well as orthographic words (word boundaries are shown as vertical lines in both panes). The upper pane shows the wavelet transform in a form of scalogram as well as eight separated scales (grey lines) ranging from segmentally influenced perturbation or microprosody (lowest scale) to utterance level phrase structure (the highest level). The potentially prominent peaks in the signal occurring during most content words are clearly visible in the scalogram as red areas, while the valleys shown in blue indicate intervals with low f_0 that might be associated with prosodic boundaries.

The time-scale analysis allows for not only locating the relevant features in the signal but also estimating their relative salience, i.e., their prominence, visible as positive local extrema (red in Fig. 2). There are several ways to estimate word prominences from a CWT. Suni et al. [22] and Vainio et al. [23] used amplitude of the word prosody scale which was 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. A more sophisticated way is presented in [24] where the lines of maximum amplitude (LoMA) in the wavelet transform of f_0 contours were used [42, 31, 43]. This method was shown to be on par with human estimated prominence values (on a four degree scale). However, the method suffered from the fact that not all prominent words are identified and – more importantly – some words are estimated as prominent whereas they should be seen as non-prominent parts of either another phonological word or a phrase.

The current study extends this approach in two ways. First, instead of using f_0 signal the CWT is performed on a composite signal combining f_0, energy and word duration information (Section 2.3). Second, estimation of word prominences is supplemented by a procedure for identification of prosodic boundaries using the same CWT approach. The strength of prosodic boundary for each word can be estimated using lines of *minimum* amplitude at word boundaries (Section 2.2).

In the remaining part of the section we describe the main steps for analysing and annotating prominences and boundaries in a fully automatic and unsupervised fashion using the CWT and LoMA on composite prosodic signal based on fundamental frequency, intensity, and timing.

2.1. Wavelet decomposition

The basis for the modeling of hierarchies in speech signals is provided by continuous wavelet transform (CWT) [44, 45]. The CWT is a decomposition of a signal to a number of scales. To define the transform, let s be a one-dimensional signal with real values and finite energy. Given a scale $\sigma > 0$ and a temporal translation τ, the continuous wavelet transform can be defined as $W_s(\sigma, \tau) = \sigma^{-1/2} s * \psi_{\tau,\sigma}$ where $*$ denotes convolution and $\psi_{\tau,\sigma}$ is the Mexican hat mother wavelet translated by τ and dilated by σ.

The Mexican hat mother wavelet belongs to a family of Gaussian wavelets. These wavelets seem to give a suitable compromise between temporal and frequency selectivity in the time-frequency representation of the prosodic signals. Although the Mexican hat mother wavelet has infinite support, the values decay exponentially fast far away from the origin and the mother wavelet effectively acts on a support of seven units.

The sampling rate of a digital signal determines the finest temporal scales available for the analysis. In the statistical speech synthesis context a 5 ms fixed window size is used for acoustical parameters. Every real signal also has finite length and the coarsest scales become obsolete. The onset and offset of the signal can create artifacts propagating to the wavelet image and here these effects are counteracted by continuing the signal periodically.

The original signal s can be reconstructed approximately from the original signal using a finite number of wavelet scales with

$$s(t) \approx c \sum_{j=0}^{N} a^{-j/2} \, W_s(a_0 a^j, t)$$

where $a_0 > 0$ is the finest (smallest) scale, $a > 1$ defines the spacing between chosen scales, $N > 1$ is the number of scales included, and c is a constant.

2.2. Lines of maximum and minimum amplitude

In order to quantify word prominence, lines of maximum amplitude (LoMA) joining the nearby peaks on subsequent scales are created and each line is assigned a strength representing the cumulative sum of scale values forming the

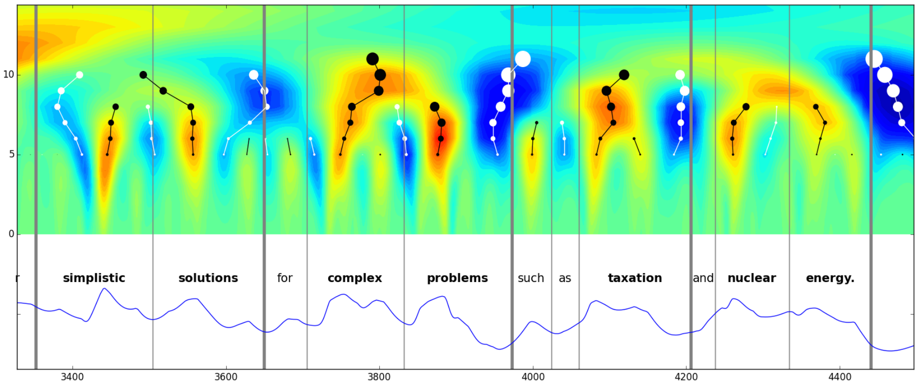

Figure 3: A fragment of speech from BURNC analysed with CWT-LoMA with composite prosodic signal (combining energy, f_0 and word duration). Maxima lines are drawn in black and minima lines in white, with point size representing cumulative strength. Annotated prosodic boundaries are marked with vertical lines and accented words with boldface type.

line (black lines in Fig. 3). This is similar to the rhythmogram approach [34, 35] which has also been applied to prominence detection [46].

Formally, LoMAs are defined recursively by connecting local maxima across scales. First, let $t_{1,0}$, $t_{2,0}$,..., $t_{n,0}$ be the time points where the local maxima occurred in the finest scale ($\sigma = a_0$) in descending amplitude order, $W_s(a_0, t_{1,0}) \geq \ldots \geq W_s(a_0, t_{n,0})$. Then the point $t_{i,0}$, $i = 1, \ldots, n$ is connected to the nearest local maximum (the mother candidate) to the right at the scale $a_0 a$ if the derivative along the scale at $t_{i,0}$ is positive, the distance to the mother candidate is at most 200 ms, and the mother candidate was not connected to a child earlier. If the derivative was negative, the search was done to the left. For consecutive levels, the ordering is based on the cumulative weighted sum of the local maximum together with its descendants: for a local maximum in $t_{i,j}$, $j > 0$, at level $a_0 a^j$, with descendants in $t_{i_0,0}, \ldots, t_{i_j,j}$ at levels $a_0, \ldots, a_0 a^j$ respectively, the cumulative weighted sum is

$$W_s(a_0, t_{i_0,0}) + \ldots + \log(j+1)\, a^{-j/2}\, W_s(a_0 a^j, t_{i_j,j}).$$

Without the logarithmic term in the above sum, the formula resembles a lot the reconstruction of the original signal. Since the local maxima often are close to each other, the logarithmic term plays a crucial role in giving more weight to the higher levels of hierarchy. Observe that the number of local maxima decrease with increasing scales, every local maximum has at most one parent, and every parent has exactly one child. Finally, the points connected as children and parents form lines of maximum amplitude (LoMA) and the *strength* of such a line is the weighted sum of all the elements included in the line.

In a corresponding manner, the lines of minimum amplitude (LomA) (shown in white in Fig. 3) used for evaluation of boundary strength join the local minima

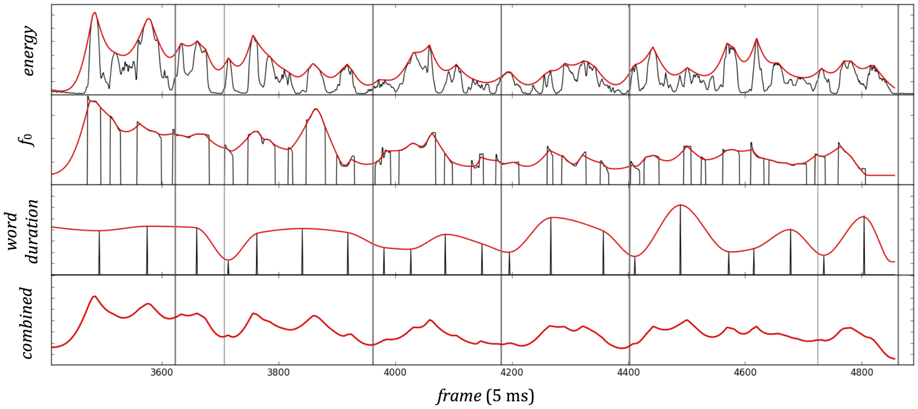

Figure 4: Prosodic parameters used in LoMA analysis extracted from BURNC. Raw parameters are drawn in gray and interpolated final parameters are shown in red. Combined prosodic signal is shown in the bottom. Gray vertical lines represent manually annotated prosodic boundaries.

of scales; formally LomAs of a signal s are defined as the lines of maximum amplitude of $-s$.

2.3. Preprocessing of the signals

The acoustic signal reflects the physiological control actions behind speech communication. Emphasised words are often louder, higher in pitch, and longer as a result of more production effort, higher fundamental frequency, and prolonged duration. For analysing the acoustic patterns, the abrupt changes in f_0 or gain, due to e.g. closures in the vocal tract during stops, create strong hierarchical structures in the wavelet image that might not be part of the auditory *gestalt* [47]. In order to better represent the more continuous underlying articulatory gestures and seemingly more continuous percepts, the acoustic signals are "filled in" for the portions where signal cannot be found (for f_0) or where it is very weak (gain). In addition, a continuous (with respect to the time) representation for duration is derived. Although inspired by the physiology of vocal and auditory apparatuses, the aim of these transformations is not to model these systems but to make the algorithm more comparable to the other phenomenological approaches to describe the key prosodic patterns.

2.3.1. Intensity

Intensity variations in the speech signal are primarily caused by (deliberate and random) fluctuations of subglottal pressure and the degree of hyperarticulation (especially in fricatives). As a proxy to the articulatory effort, the gain of the acoustical signal is transformed by iteratively interpolating the silent gaps.

Let ϕ be the Gaussian kernel and g the original gain signal (i.e. a logarithm of the amplitude). A family of scaling functions, $\{\phi_i\}_i$ is obtained by dilating and scaling ϕ with constants $\lambda_i = w_{\text{max}}^{(i-n)/n} w_{\text{min}}^{-i/n}$, $i = 0, 1, 2, \ldots, n$, where w_{max}

is the maximum smoothing window size, $w_{\min}$ is the minimum window, and n is the size of the family. The g is recursively smoothed. For $i = 0$, a pointwise maximum is taken by $g_0 = \max\{g, g * \phi_0\}$ where $*$ denotes the convolution. For $i > 0$, $g_i = \max\{g, g_{i-1} * \phi_i\}$. This results in the preprocessed gain $g = g_n$ shown in the top panel of Figure 4.

2.3.2. Fundamental frequency

The auditory pitch of the voiced sounds is closely related to the lowest eigen resonances of the vocal folds. However, during unvoiced speech segments, the association between the acoustic signal and the eigen resonances of the vocal folds break apart. Importantly, even during the silent periods there are control actions to the vocal folds that impact the f_0 once the vibration its reinitiated either by adducting the vocal folds or by restoring the airflow through the vocal tract. In addition to the internal state of the larynx, the frequency of the glottal pulsing is influenced by the subglottal pressure. Not surprisingly then, the f_0 and intensity are strongly correlated. To estimate the state of the f_0 control during unvoiced portions, an algorithm is proposed where the surface f_0 values are left unchanged for the voiced passages and the underlying state of the vocal folds is estimated by interpolation for unvoiced passages.

The gap filling for the unvoiced portions of fundamental frequency signal s is similar to that for the gain. First, the signal is decomposed in voiced and unvoiced portions by defining the set V of time points where the speech signal is voiced.

In practice, the voicedness of a time point is defined using the GlottHMM [48] analysis which applies low-frequency energy and zero-crossings thresholds for voicing decision. Then, using the same smoothing family as before, the smoothed s is defined iteratively: for $i = 0$, $s_0 = s\chi_V + \max\{s, s*\phi_0\}\chi_{V^C}$ where χ_A is the characteristic function of a set A and A^C denotes the complement of the set A. The analogous recursive formula then is

$$s_i = s\chi_V + \max\{s, s_{i-1} * \phi_i\}\chi_{V^C}$$

resulting in the preprocessed fundamental frequency. Finally, to remove perturbation around gaps, the iterated signal s_n is smoothed using the same iterated maximisation algorithm as for the gain.

To find suitable parameters in the above algorithms, two test utterances were used. The following values were used: $w_{\max} = 100$ ms, $w_{\min} = 1$, for both gain and f_0; $n = 100$ for for gain, $n = 200$ for f_0. For the final smoothing of f_0 in order to remove perturbation around gaps (see above): $w_{\max} = 25$ ms and $n = 50$.

Observe that the repeated convolutions and maximums do not let the signals grow in an unlimited way. Instead, every point converges and the resulting (maximal) function has comparable energy to the original which can be seen by iterating a result of Hardy and Littlewood [49], (for modern approach, see Theorem 2.19 in [50]).

2.3.3. Duration

The duration of a phonological unit varies as a function of its position within an utterance. For instance the speech rate often changes across boundaries and accented words are longer. Due to a lack of signal based speech rate estimators, the duration signal has to be based on analytical linguistic units rather than the raw signal. To quantify the duration, a relation between acoustical (continuous) duration and a suitable discrete linguistic unit is needed. A natural candidate could be a syllable but here an orthographic word is chosen instead as the syllable boundaries might not be easy to derive from text without supervision. To apply the wavelet analysis to the duration, it is expanded to a continuous time dependent variable which ideally would reflect the local duration of the linguistic units. For the current experiment provided word alignments were used. The word boundaries, $x_0, x_1, \ldots, x_{N_w}$, where N_w is the number of orthographic words within a given speech signal, and the associated durations $d_i = x_i - x_{i-1}$, $i = 1, \ldots, N_w$, are computed. The points $\{(x_{i-1} + d_i/2, d_i)\}$ are connected using cubic splines to yield a duration signal d defined for every time instant from x_0 to x_{N_w} with the same sampling rate as for fundamental frequency and gain (see third panel from the top in Fig. 4). When annotated pauses and breaths between words occurred, these were not taken into consideration, i.e. the duration of these gaps was ignored by the interpolation procedure.

To approximate the local speech rate, the time derivative of the duration signal d was used instead of the continuous duration.

2.4. Annotation

The annotation of accents and breaks (prominences and boundaries) is based on wavelet decomposition of the fundamental frequency, gain, and duration derivative signals. These three acoustic signals were normalized to have unit variance and then different combinations of these signals (see Section 3.2) were summed to yield the evaluated prosodic signal s.

To normalize the speech rate, the finest scale was selected for each utterance separately through finding the word scale a_W which is the ratio of word count and utterance duration. For word prominence evaluation, the finest scale used was one octave below the word scale, i.e. $a_0 = a_W/2$. As prosodic breaks manifest mostly on larger scales, the word scale was taken as the finest scale $a_0 = a_W$ for boundary annotation. For both annotations, the coarsest scale was three octaves above the finest scale, i.e. $8a_0$.

3. Experimental Evaluation

As stated in the introduction, a solid method for annotating prosody would be very welcome in the field of speech synthesis, where recent development has concentrated on the acoustic modelling [51]. The motivation is crucial in building speech synthesizers for low-resourced languages, where neither linguistically nor prosodically annotated corpora are available [3]. In this chapter, we asses the utility of the proposed CWT-LoMA representation of prosody on the tasks of unsupervised annotation of prosodic prominences and boundaries. Although this hierarchical method lends itself naturally to multi-level prosody annotation [23], here, we restrict ourselves to binary detection task, in order to produce comparable results with previous studies.

Previous work on unsupervised prosody annotation has focused on accent or prominence. For example, Ananthakrishnan & Narayanan [52] performed two-class unsupervised clustering on syllable level acoustic features combined with lexical and syntactic features, achieving accent detection accuracy of 78 % using the Boston University Radio News Corpus (BURNC). In a similar vein, Mehrabani et al. [53] annotated a corpus with four level prominence scale by K-means clustering on foot-level acoustic features, achieving improved synthesis quality compared to a rule-based prominence model. Using a more analytic approach, Tamburini [54] derived a continuous prominence function, using expert knowledge to weight various acoustic correlates of prominence, achieving 80 % accuracy on syllable prominence detection on the TIMIT corpus. Word prominence was annotated by Vainio & Suni [55] with a similar approach, additionally using the differences between synthesized and original prosodic features as a normalizing method. An ambitious approach was presented by Kalinli & Narayanan [56], who extracted multi-scale auditory features insipired on the processing stages in the human auditory system, combined to an auditory salience map. They achieved prominent word detection accuracy of 78 % with F-score of 0.82 on BURNC, which, to our knowledge, is the best reported unsupervised result on this corpus to date. Similar to the method proposed here, Ludusan [46] applied a hierarchical rhythmogram model of loudness to annotate syllable prominence, achieving significant improvement over raw prosodic features in terms of accuracy on multiple corpora.

Whereas the text-based break prediction studies are abundant due to its importance in TTS, unsupervised acoustic boundary estimation has received less interest. This probably stems from the fact that both acoustic pauses, which can be obtained reliably by HMM forced alignment, and punctuation yield high baseline accuracy on major boundaries, and for TTS purposes, this has been considered satisfactory. For example in BURNC, intonational phrase boundaries can be predicted by silence alone with 88 % accuracy, though with only 45 % recall, and traditional acoustic features offer little improvement over this trivial baseline [57]. In terms of combining text and acoustic evidence, Ananthakrishnan & Narayanan [52] obtained 81 % accuracy in combined intermediate and intonational boundary detection with a two class k-means model.

3.1. Corpus

We performed the evaluation of our prominence and boundary detection method on Boston Radio News corpus [58], chosen for its high quality prosodic labeling and comparability with several previous methods also evaluated on BURNC. The corpus consists of about two and a half hours of news stories read by 6 speakers with manual Tone and Break Index annotations. The ToBI labelling scheme was originally developed for transcribing phonologically distinctive elements of speech melody [2], thus high (H), low (L) and complex accent types are employed (H*, L*, L*+H, L+H*, H+ !H*), denoting syllable level shape and peak alignment. Prosodic boundaries are annotated with boundary tones (L-, H-,L – L%, L – H%, H – H%, H – L%), again signalling melodic shapes. Break strength is annotated in the form of break indices ranging from zero (clitized) to four (intonational phrase boundary). For the boundary detection task, we considered a word boundary as a prosodic boundary if the last syllable of a preceding word was marked with break index three (intermediate phrase break) or four (intonational phrase break). Prominence, on the other hand, has not been directly annotated and for the current experiment, we made a simplifying assumption that a word is prominent if any of its syllables carries an accent. These binary boundary and prominence categories are consistent with previous prosodic event detection studies [59, 60]. Almost all of the annotated data were used for the experiment, totalling 442 stories or 29774 words. Three stories from speaker f2b, used for setting values of free parameters were excluded as well as few cases were syllable and word alignments did not match. Word level break and prominence labels were derived by combining the provided, time aligned syllable and word labels. Manually corrected alignments were used when available.

3.2. Features and Processing

The proposed method was evaluated using standard prosodic features; f_0, energy and word duration, as well as all combinations of those. Raw f_0 and energy parameters were analyzed from 16 kHz speech signals with GlottHMM analysis-synthesis framework [48] with five millisecond frame shift. The method uses iterative-adaptive inverse filtering to separate the contributions of vocal tract and voice source, and performs f_0 analysis on the source signal with autocorrelation method. Log energy is calculated from the whole signal. Pitch range was set separately for male and female speakers, 70–300 Hz and 120–400 Hz, respectively. The obtained f_0 and energy parameter tracks were interpolated using a peak preserving method. Word durations were transformed to continuous signals as described in Section 2.3. Labeled pauses and breaths were not considered in the duration transform. When evaluating the performance of combinations of prosodic features, the individual parameters were normalized utterance-wise to zero mean, unity variance, and summed prior to the wavelet analysis, after which the composite prosodic signal was again normalized.

The signal was then used as such in the wavelet analysis, without any feature extraction step. Continuous wavelet transform was performed using the second derivative of gaussian (Mexican hat) wavelet, with a half octave scale separation. A scale corresponding to word level was estimated individually for each paragraph in order to normalise the differences caused varying speech rate. Lines of maximum and minimum amplitude were then estimated from the scalogram. The strongest peak LoMA of each word was assigned as the prominence value of the word and strongest valley LomA between each two word's strongest peak LoMA as a boundary value. If a word contained no peak LoMA, valley LomA was searched between the midpoints of adjacent words. Further, if either peak LoMA or valley LomA was not found, prominence or boundary value was set to zero respectively. To verify the utility of hierarchical modelling and to rule out the possibility that improvements were achieved only due to feature engineering, we also calculated word maximum (to represent prominence) and minimum between midpoints of adjacent words (to represent boundary) from the raw prosodic signal to be used as a baseline. In order to compare the predicted continuous prominence and boundary values against manual labels, the values were converted to a binary form by searching for an optimal value for separating the two classes in terms of classification accuracy, using 10 % of the manual labels. Although continuous values could be used as such in other applications, it might be argued that this step weakens our claim for unsupervision for the current task. Thus, for the best configurations, we also report results based on dividing the prominence and boundary distributions to two classes by unsupervised k-means clustering.

3.3. Results

Results on CWT-LoMA analysis of f_0 (f0) energy (en) and duration (dur) and their combinations on prominence and boundary detection are reported here. The performance of gap-filling on energy is evaluated separately, and whenever the gap-filling improves the performance for prominence or boundary annotation, it is used for energy in combined features as well. Boundaries were defined as manual break indices of either 3 or 4; prominence if any syllable of a word carries an accent. Results are presented on word level, in terms of percentage of correct detections, i.e. accuracy, as well as precision, recall and F-score. As baselines, we report the majority class, predictions derived from the best combination signal without wavelet analysis, as well as current state-of-the-art unsupervised and supervised acoustic results. Note that these results are only roughly comparable, as there are minor differences in data selection. The results presented in Table 1 show an improvement of both prominence and boundary detection compared to baseline unsupervised methods when the combined signal is used. More importantly, the CWT method further improves the detection in terms of accuracy, F-score, precision and recall. Interestingly, the improvement in accuracy over non-CWT approach (*f0_en_dur_raw*) is significant ($p < 0.001$) not only for the CWT method using all three signal components (*f0_en_dur*) but also for several other combinations of prosodic signals.

Strictly unsupervised results using two class k-means clustering on the prediction distributions using all acoustic features were 84.0 % accuracy and 0.86 F-score for prominence and 85.5 %, 0.73 for boundary detection respectively.

Examining the results of individual acoustic features, we note similar perfomance for f_0 and energy in both tasks, and word duration not far behind. f_0 appears more important for prominence detection, which is expected as the reference labeling concerned pitch accents. Filling the unvoiced gaps of the energy signal helps in boundary detection, but not in the accent detection task, perhaps due to syllable level features of the signal being smoothed too much. Interestingly, combining f_0 and energy yields only modest improvement, whereas combining either with duration provides substantial gain; accuracy increases approximately 3 % in accent detection and 4 % in boundary detection. Though

Table 1: Summary of results on BURNC with comparison to earlier experiments. The bolded figures depict the best results both in current experiments and the literature. [1]Kalinli & Narayanan [56],[2]Rosenberg & Hirchberg [61],[3]Ananthakrishnan & al. [52],[4]Ananthakrishnan & al. [59]. Asterisks mark the significance of accuracy % being higher for the methods using CWT analysis in comparison with non-hierarchical *f0_en_dur_raw* baseline (Wilcoxon signed rank test with continuity correction $^{***} < 0.001 < {}^{*} < 0.05$)

	Prominence Detection				**Boundary Detection**			
feature	*acc.%*	*F-score*	*prec.*	*rec.*	*acc.%*	*F-score*	*prec.*	*rec.*
CWT-method								
f0	80.9***	0.82	0.84	0.81	81.1	0.56	0.79	0.44
en	79.5*	0.83	0.77	0.89	78.6	0.54	0.68	0.45
en_interp.	78.3	0.81	0.77	0.86	81.0	0.56	0.79	0.44
dur	79.5	0.81	0.81	0.81	80.3	0.64	0.66	0.61
f0_en	82.5***	0.85	0.81	0.88	81.7	0.59	0.80	0.47
f0_dur	84.2***	**0.86**	0.85	0.87	**85.7*****	0.72	0.79	0.67
en_dur	82.5***	0.84	0.82	0.86	85.2***	**0.73**	0.75	0.70
f0_en_dur	**84.6*****	**0.86**	0.84	0.90	**85.7*****	0.72	0.80	0.65
Baselines								
majority	54.5				72.0			
f0_en_dur_raw	79.2	0.81	0.82	0.80	82.1	0.62	0.76	0.53
unsupervised[1,2]	78.1	0.82	0.78	0.86	81.1	0.66	0.64	0.69
acoustic sup.[3,4]	84.2	**0.86**			84.6			

a naïve feature, word duration may capture both lengthening effects as well as lexical information, separating most of the function and content words, and disambiguating the alignment of LoMA. Combining all features provides best results, but not by a significant margin. Comparison of the detection estimates from raw combined signal to ones provided by CWT-LoMA confirm the importance of hierarchical modelling with a solid advantage in both tasks.

Compared to previous methods, our results improve upon the unsupervised state-of-the-art by a significant margin, and at least match the accuracy of acoustic-based supervised methods. The results are not far from performance of supervised methods using acoustic, lexical, and syntactic evidence, where reported accuracies for both word level prominence and boundary detection range from 84 % to 87 % [59, 60].

4. Discussion and Conclusions

The results show that prominences and boundaries can be viewed as manifestations of the same underlying speech production process. This has, of course, many theoretical implications. As foremost is the fact that the suprasegmental variables used (f_0, energy envelope, duration) seem to work seamlessly to the same end, which is to signal the hierarchical and parallel structure of the linguistic signals. The role of signal energy as a reliable determinant of prosodic structure is interesting, but not altogether surprising [62]. On the one hand, it diminishes the role of f_0, while on the other hand, it also provides it with more freedom for other (post-lexical) prosodic functions that are not strictly related to the hierarchical structure.

In contrast to most published work on speech prosody, the results here show that prosodic structure can – and probably should – be studied and represented in a unified framework comprising all relevant signal variables at the same time. Although our methodology is unsupervised and purely signal based and does not model – explicitly or implicitly – the top-down processes associated with prosodic parsing by humans, the system models the judgements of human annotators to a considerable degree. This suggests that at least some of the top-down aspects of prominence and boundary identification are sufficiently captured by the hierarchical nature of our analysis that takes into consideration contextual information present in the signal on different time scales.

Prominence (and, to a lesser degree, boundaries) may also be marked without the increase in intensity/pitch/duration that is implicitly assumed in our approach or, indeed, by intentional violation of this expectation. While using a combination of all three signal dimensions simultaneously facilitates detecting prominent units with at least one of these attributes magnified, our method does not fully supplant the human-like top-down processing that might be needed to detect these cases. In fact, the method presented here can be used to automatically identify this type of phenomena (in a labeled corpus) for further, more detailed study. Also, languages differ both in terms of their hierarchical structure and the way their speakers mark prominence. One of the means to address this language-dependency using the present approach is to vary the way the signal dimensions (duration, intensity, f_0) are weighted in the combined signal. The future empirical work on this issue is necessary to establish to what extent our methodology can be applied to other languages.

As mentioned above, the methods and representations brought forward in this study have been designed to be feasible in a broader scientific spectrum keeping in mind their psychological plausibility. Although the wavelet representation of prosody has a strong correspondence with the manual annotations of the evaluation corpus (highlighting their relationship with perception), the neural computations performed by the auditory system might differ considerably in contributing to the percepts underlying the accent and break judgments of the labellers. In particular, the scheme for iteratively filling the gaps in the acoustic signals is not a plausible algorithm for neural processing. However, the assumed temporal integration model [63, 64] to explain silent gap detection gives similar "filling" behaviour as the current processing of gain signal. Importantly, the parameters and particular formulas were only inspired by the known auditory processes but chosen based on performance on a few test sentences. In the future, more principled grounding of our approach in perceptual processes, e.g., akin to auditory primal sketch theory of Todd [34, 35], should be investigated.

Also, in the proposed accent and boundary annotation the wavelet analysis is performed to a few one-dimensional signals. However, a neurally more plausible approach would be a truly multidimensional representation of speech signal similar to the multi-scale visual analyses [26]. Crucially, the wavelet trees relate the accents and boundaries together phonetically hinting at a unified mechanism, in both production and perception, between the phonetic realisation of these primary concepts of prosodic phonology.

Acknowledgements

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 (Simple4All) and the Academy of Finland (project n^{o} 1265610 (the MIND programme)) as well as project n^{o} 293346 (the Digital Humanities programme). We are indebted to Petra Wagner and two anonymous reviewers for their insightful comments and suggestions.

References

[1] C.-Y. Tseng, S.-H. Pin, Y. Lee, H.-M. Wang, Y.-C. Chen, Fluent speech prosody: Framework and modeling, Speech Communication 46 (3) (2005) 284–309.

[2] K. E. A. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, J. Hirschberg, ToBI: A standard for labeling English prosody., in: ICSLP, Vol. 2, ISCA, 1992, pp. 867–870.

[3] Simple4All speech synthesis project (EU-FP7) (2014) [cited February 15, 2015].
URL http://www.simple4all.org

[4] J. A. Goldsmith, Autosegmental and metrical phonology, Vol. 11, Blackwell Oxford, 1990.

[5] M. Halle, J.-R. Vergnaud, et al., Metrical structures in phonology, Ms. Cambridge, MA.

[6] M. Halle, J.-R. Vergnaud, Three dimensional phonology, Journal of linguistic research 1 (1) (1980) 83–105.

[7] D. H. Klatt, Review of text-to-speech conversion for english, The Journal of the Acoustical Society of America 82 (3) (1987) 737–793.

[8] S. Öhman, Word and sentence intonation: A quantitative model, Speech Transmission Laboratory, Department of Speech Communication, Royal Institute of Technology, 1967.

[9] H. Fujisaki, H. Sudo, A generative model for the prosody of connected speech in Japanese, Annual Report of Engineering Research Institute 30 (1971) 75–80.

[10] H. Fujisaki, K. Hirose, Analysis of voice fundamental frequency contours for declarative sentences of Japanese, Journal of the Acoustical Society of Japan (E) 5 (4) (1984) 233–241.

[11] G. Bailly, B. Holm, Sfc: a trainable prosodic model, Speech Communication 46 (3) (2005) 348–364.

[12] G. K. Anumanchipalli, L. C. Oliveira, A. W. Black, A statistical phrase/accent model for intonation modeling., in: INTERSPEECH, 2011, pp. 1813–1816.

[13] G. Kochanski, C. Shih, Stem-ml: language-independent prosody description., in: INTERSPEECH, 2000, pp. 239–242.

[14] G. Kochanski, C. Shih, Prosody modeling with soft templates, Speech Communication 39 (3) (2003) 311–352.

[15] A. Eriksson, G. C. Thunberg, H. Traunmüller, Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing, in: Proc. European Conf. on Speech Communication and Technology Aalborg, September 2001, Vol. 1, 2001, pp. 399–402.

[16] J. Cole, Y. Mo, M. Hasegawa-Johnson, Signal-based and expectation-based factors in the perception of prosodic prominence, Laboratory Phonology 1 (2) (2010) 425–452.

[17] D. Arnold, P. Wagner, B. Möbius, Obtaining prominence judgments from naïve listeners–influence of rating scales, linguistic levels and normalisation, Proceedings of Interspeech 2012 (2012) 2394–2397.

[18] M. Vainio, A. Suni, T. Raitio, J. Nurminen, J. Järvikivi, P. Alku, New method for delexicalization and its application to prosodic tagging for text-to-speech synthesis, in: Interspeech, Brighton, UK, 2009, pp. 1703–1706.

[19] M. Vainio, J. Järvikivi, Tonal features, intensity, and word order in the perception of prominence, Journal of Phonetics 34 (2006) 319 – 342.

[20] A. Eriksson, E. Grabe, H. Traunmüller, Perception of syllable prominence by listeners with and without competence in the tested language, in: Speech Prosody 2002, International Conference, 2002, pp. 275–278.

[21] P. Wagner, A. Origlia, C. Avesani, G. Christodoulides, F. Cutugno, M. D'Imperio, D. Escudero Mancebo, B. Gili Fivela, A. Lacheret, B. Ludusan, H. Moniz, A. Ní Chasaide, O. Niebuhr, L. Rousier-Vercruyssen, A.-C. Simon, J. Šimko, F. Tesser, M. Vainio, Different parts of the same elephant: a roadmap to disentangle and connect different perspectives on prosodic prominence, in: Proceedings of the 18th International Congress of Phonetic Sciences, 2015.

[22] A. Suni, D. Aalto, T. Raitio, P. Alku, M. Vainio, Wavelets for intonation modeling in HMM speech synthesis, in: 8th ISCA Speech Synthesis Workshop (SSW8), Barcelona, Spain, 2013, pp. 285–290.

[23] M. Vainio, A. Suni, D. Aalto, Continuous wavelet transform for analysis of speech prosody, TRASP 2013 - Tools and Resources for the Analysis of Speech Prosody, An Interspeech 2013 satellite event, August 30, 2013, Laboratoire Parole et Language, Aix-en-Provence, France, Proceedings (2013) 78–81.

[24] M. Vainio, A. Suni, D. Aalto, Emphasis, word prominence, and continuous wavelet transform in the control of hmm-based synthesis, in: Hirose, Tao (Eds.), Speech Prosody in Speech Synthesis - Modeling, realizing, converting prosody for high quality and flexible speech synthesis, Springer, 2015, pp. 173–188.

[25] J. C. Russ, R. P. Woods, The image processing handbook, Journal of Computer Assisted Tomography 19 (6) (1995) 979–981.

[26] B. M. ter Haar Romeny, A geometric model for the functional circuits of the visual front-end, in: Brain-Inspired Computing, Springer, 2014, pp. 35–50.

[27] G. Zweig, Basilar membrane motion, in: Cold Spring Harbor Symposia on Quantitative Biology, Vol. 40, Cold Spring Harbor Laboratory Press, 1976, pp. 619–633.

[28] T. Altosaar, M. Karjalainen, Event-based multiple-resolution analysis of speech signals, Proc. of IEEE ICASSP-88. New York (1988) 327–330.

[29] X. Yang, K. Wang, S. A. Shamma, Auditory representations of acoustic signals, Information Theory, IEEE Transactions on 38 (2) (1992) 824–839.

[30] R. Ramachandran, R. Mammone, Modern methods of speech processing, Springer, 1995.

[31] M. D. Riley, Speech Time-Frequency Representation, Vol. 63, Springer, 1989.

[32] H. Reimann, Signal processing in the cochlea: The structure equations, J. Math. Neuroscience 1 (5) (2011) 1–50.

[33] A.-L. Giraud, D. Poeppel, Cortical oscillations and speech processing: emerging computational principles and operations, Nature neuroscience 15 (4) (2012) 511–517.

[34] N. P. M. Todd, G. J. Brown, A computational model of prosody perception, in: Proceedings of the International Conference on Spoke Language Processing (ICLSP-94, 1994, pp. 18–22.

[35] N. P. M. Todd, The auditory primal sketch: A multiscale model of rhythmic grouping, Journal of New Music Research 23 (1) (1994) 25–70.

[36] M. H. Farouk, Application of Wavelets in Speech Processing, Springer, 2014.

[37] H. Kruschke, M. Lenz, Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis., in: Eighth European Conference on Speech Communication and Technology, 2003, pp. 2881–2884.

[38] T. Mishra, J. Van Santen, E. Klabbers, Decomposition of pitch curves in the general superpositional intonation model, Speech Prosody, Dresden, Germany.

[39] J. P. v. Santen, T. Mishra, E. Klabbers, Estimating phrase curves in the general superpositional intonation model, in: Fifth ISCA Workshop on Speech Synthesis, 2004.

[40] M. Lei, Y.-J. Wu, F. K. Soong, Z.-H. Ling, L.-R. Dai, A hierarchical f0 modeling method for HMM-based speech synthesis., in: INTERSPEECH, 2010, pp. 2170–2173.

[41] H. Zen, N. Braunschweiler, Context-dependent additive log f_0 model for HMM-based speech synthesis., in: INTERSPEECH, 2009, pp. 2091–2094.

[42] S. Mallat, A wavelet tour of signal processing, Access Online via Elsevier, 1999.

[43] A. Grossman, J. Morlet, Decomposition of functions into wavelets of constant shape, and related transforms, Mathematics and Physics: Lectures on Recent Results 11 (1985) 135–165.

[44] I. Daubechies, et al., Ten lectures on wavelets, Vol. 61, SIAM, 1992.

[45] C. Torrence, G. P. Compo, A practical guide to wavelet analysis, Bulletin of the American Meteorological society 79 (1) (1998) 61–78.

[46] B. Ludusan, A. Origlia, F. Cutugno, On the use of the rhythmogram for automatic syllabic prominence detection, in: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, 2011, pp. 2413–2416.

[47] J. Barnes, A. Brugos, N. Veilleux, S. Shattuck-Hufnagel, Voiceless intervals and perceptual completion in f0 contours: Evidence from scaling perception in american english, Proc. 16th ICPhS, Hong Kong, China (2011) 108–111.

[48] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, P. Alku, Hmm-based speech synthesis utilizing glottal inverse filtering, Audio, Speech, and Language Processing, IEEE Transactions on 19 (1) (2011) 153–165.

[49] G. H. Hardy, J. E. Littlewood, A maximal theorem with function-theoretic applications, Acta Mathematica 54 (1) (1930) 81–116.

[50] P. Mattila, Geometry of sets and measures in Euclidean spaces: fractals and rectifiability, no. 44, Cambridge University Press, 1999.

[51] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, K. Tokuda, The HMM-based speech synthesis system (HTS) version 2.0, in: Proc. of Sixth ISCA Workshop on Speech Synthesis, 2007, pp. 294–299.

[52] S. Ananthakrishnan, S. S. Narayanan, Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling, in: Proceedings of InterSpeech, Pittsburgh, PA, 2006, pp. 297–300.

[53] M. Mehrabani, T. Mishra, A. Conkie, Unsupervised prominence prediction for speech synthesis, in: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013, 2013, pp. 1559–1563.

[54] F. Tamburini, C. Caini, An automatic system for detecting prosodic prominence in american english continuous speech, International Journal of Speech Technology 8 (1) (2005) 33–44.

[55] M. Vainio, A. Suni, P. Sirjola, Accent and prominence in Finnish speech synthesis, in: G. Kokkinakis, N. Fakotakis, E. Dermatos, R. Potapova (Eds.), Proceedings of the 10th International Conference on Speech and Computer (Specom 2005), University of Patras, Greece, 2005, pp. 309–312.

[56] O. Kalinli, S. S. Narayanan, A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech, in: INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007, 2007, pp. 1941–1944.

[57] A. Rosenberg, Automatic detection and classification of prosodic events, Ph.D. thesis, Columbia University (2009).

[58] M. Ostendorf, P. Price, S. Shattuck-Hufnagel, The Boston University Radio News Corpus, Tech. rep. (2005).

[59] S. Ananthakrishnan, S. Narayanan, Automatic prosodic event detection using acoustic, lexical, and syntactic evidence, Audio, Speech, and Language Processing, IEEE Transactions on 16 (1) (2008) 216–228.

[60] O. Kalinli, S. S. Narayanan, IEEE Transactions on Audio, Speech and Language Processing 17 (5) (2009) 1009–1024.

[61] A. Rosenberg, J. Hirschberg, Detecting pitch accents at the word, syllable and vowel level, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, NAACL-Short '09, Association for Computational Linguistics, 2009, pp. 81–84.

[62] G. Kochanski, E. Grabe, J. Coleman, B. Rosner, Loudness predicts prominence: Fundamental frequency lends little, The Journal of the Acoustical Society of America 118 (2) (2005) 1038–1054.

[63] M. Penner, Detection of temporal gaps in noise as a measure of the decay of auditory sensation, The Journal of the Acoustical Society of America 61 (2) (1977) 552–557.

[64] B. R. Glasberg, B. C. Moore, S. P. Bacon, Gap detection and masking in hearing-impaired and normal-hearing subjects, The Journal of the Acoustical Society of America 81 (5) (1987) 1546–1556.