Speech Prosody 2016
31 May - 3 June 2016, Boston, USA

Boundary detection using Continuous Wavelet Analysis

Antti Suni[1,2], *Juraj Šimko*[1], *Martti Vainio*[1]

[1]University of Helsinki, Finland, [2]Aalto University, Finland
{antti.suni,juraj.simko,martti.vainio}@helsinki.fi

Abstract

Unsupervised boundary detection and classification is both a theoretically interesting question and an important challenge for speech technology. Theoretical interest lies in exploring how and to what extent is the boundary information encoded in purely acoustic material. For technology, automatic boundary detection facilitates cheap and fast labeling of large corpora of speech data. In this work we present a novel methodology of automatic and unsupervised boundary detection and classification based on the continuous wavelet transform (CWT) technique. Several approaches using lines of minimal amplitude, phase information and wavelet-based estimation of speech tempo are evaluated and compared on Boston Radio News Corpus data. The results show that this methodology using hierarchical information encoded in speech signal compares favorably with traditionally used supervised boundary detection techniques using acoustic information.

Index Terms: boundary detection, continuous wavelet transform, speech synthesis

1. Introduction

One of the most primary features of speech prosody have to do with chunking speech into linguistically relevant units. Boundaries of various strength give rise to a hierarchy of speech constituents. In this paper we present a novel method of boundary detection based on an analysis of acoustic signal (fundamental frequency and energy). We use a continuous wavelet transform (CWT), which is in itself a hierarchical signal processing technique. The work primarily stems from a requirement to annotate speech corpora automatically, in an unsupervised fashion for text-to-speech synthesis (TTS) [1]. However, the presented representations should be of interest to anyone working on speech prosody.

Multiple methods of boundary detection have been proposed. In text-to-speech synthesis framework, it is a common practice to use punctuation and detected silences to represent phrase boundaries. This simple method achieves good precision, but yields poor recall, missing most of the more subtle boundaries (see e.g. [2]).

To cope with the finer boundary types, more refined methods aimed at improving detection recall have been proposed. In the context of the present work, these methods can be distinguished along two dimensions: to supervised and unsupervised techniques, and by the type of parameters and features they make use of (acoustic, linguistic, etc.).

Supervised methods use machine learning, data-driven approach to identify appropriate local characteristics of acoustic and other (lexical, syntactic) features associated with presence of prosodic boundary [3]. While these techniques achieve good accuracy in boundary prediction, they require considerable prior input of skilled specialists in annotating large volumes of speech corpora used as a training material.

To avoid substantial expense associated with data labeling, unsupervised methods have been also explored aiming to locate prosodic boundaries using features that can be extracted directly from speech waveform (such as f_0 or gain). Although this effort has been relatively successful, the existing techniques (e.g., [4]) nevertheless make use of some annotated features, predominantly associated with speaking rate variations known to be associated with prosodic boundaries (phrase-final lengthening). The state-of-the-art unsupervised labeling approach [4], for example, makes use of syllable nuclei durations that need to be extracted from speech material using a speech recognition system, that is, a supervised method. While recognition systems such as force aligners are widely available for most well-researched languages, they might not exist for less resourced ones.

In this work we present a novel unsupervised method of boundary detection exploiting hierarchically organized speech information as revealed by wavelet analysis. The method aim at using purely those prosodic features that can be extracted from speech waveform using speech processing techniques. We thus explore a method that in addition to f_0 and gain also extracts tempo information directly from waveform using a CWT technique. We compare its accuracy in prediction with a wavelet-based method using directly word durations from the annotated corpus. Furthermore, we evaluate two procedures of boundary prediction, one exploiting full information provided by wavelet analysis while the other relying on phase-reset only. Finally, we offer comparison to previous unsupervised and supervised results on Boston Radio News Corpus (BURNC) [5].

1.1. Continuous Wavelet Analysis

Time-scale representation based on Continuous Wavelet Analysis (CWT) presents a natural way of depicting a hierarchical prosodic structure of a complex signal, such as speech. This technique emerged independently in physics, mathematics, and engineering, and is currently widely used for analysis of complex signals including electrophysiological, visual and acoustic signals [6]. The wavelets have found applications in several areas related to speech prosody, such as robust speech enhancement in noisy signals, automatic speech segmentation, and segregation along various dimensions of speech signal in a similar way as mel-cepstral coefficients [7, 8, 9].

The advantage of using this type of analysis on a range of temporal scales is that the inherent hierarchical nature of the signal becomes visible. The analysis not only shows how information is distributed in time, but also reveals the possible interdependencies between the hierarchical levels. This property has been successfully used for predicting word prominence [10] using wavelet transform of f_0 signal. Recently, CWT de-

10.21437/SpeechProsody.2016-55

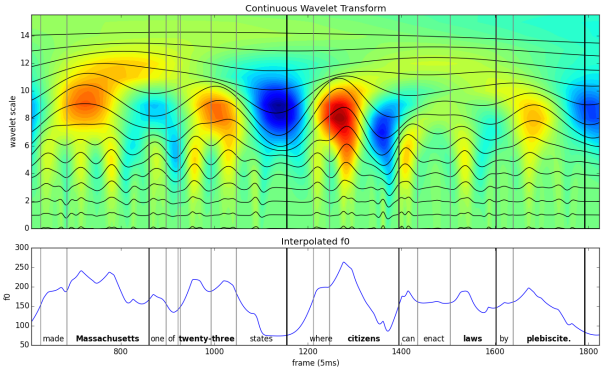

Figure 1: CWT analysis of an f_0 contour of a phrase from BURNC corpus (bottom panel). The red areas correspond to prominent portions of speech while low (blue) areas indicate a possible presence of a prosodic boundary.

composition of f_0 contour has been used to train a parametric statistical speech synthesis system [11]. Both objective and perceptual evaluation of this method explicitly targeting the inherent hierarchical nature of prosody shows an improvement in f_0 prediction and synthesis quality [12].

Fig. 1 shows a scalogram obtained by CWT of an f_0 contour of an English sentence. 14 separated scales separated by half an octave are superimposed over the scalogram heat-map. The scales can be associated with various levels of prosodic hierarchy, isolating syllable, word, phrase and utterance level contributions of the prosodic signal. Peaks and valleys at each scale correspond to predominant f_0 contour shape reflecting a given hierarchical level. Like any signal, each scale can be represented in terms of its instantaneous *amplitude* and instantaneous *phase* which can be at any given time combined to the instantaneous scale value.

2. Methods

2.1. Extraction of parameters, signal conversion

In this work we use four continuous prosodic features derived from speech signal: fundamental frequency f_0, gain (energy), continuous durational parameter (derived from labeled word durations) and an instantaneous speaking rate signal derived directly from the waveform using CWT.

Raw f_0 and energy parameters were extracted by Glott-HMM analysis-synthesis framework [13] using Iterative-adaptive inverse filtering to separate the contributions of vocal tract and voice source, performing 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. Obtained f_0 and energy parameters were interpolated using a peak preserving method (see [14] for details).

Word durations, annotated in the analyzed BURNC corpus, were first transformed to a continuous signal by cubic spline interpolation of word duration values placed at mid-points of associated word time-intervals. A time-derivative of this signal was used as word duration parameter.

One of the objectives of this paper is to evaluate a new method of estimating speaking rate using CWT analysis. The existing methods typically estimate the speech rate based on explicit identification of syllable nuclei or boundaries. Thus, the research has concentrated on signal representations where the individual syllables can be robustly identified, for example by identifying suitable spectral sub-bands and their correlations [15] or using group delay function [16].

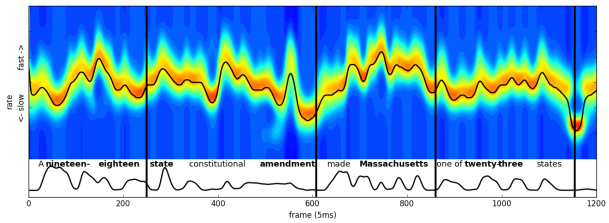

Figure 2: Illustration of speaking rate extraction method. The energy scalogram has been normalized per each frame for visual clarity (no normalization is necessary in the method). The black curve in the upper panel is the extracted rate signal, the curve in the lower panel is the processed energy envelope.

In contrast, the proposed method does not attempt to identify individual syllables, but simply extracts the temporal progression of the dominant frequency in a given prosodic signal. The speech envelope is considered to have a quasi-periodic structure with alternating peaks and valleys associated of voiced syllable nuclei and consonantal intervals. Because this alternating pattern is far from isochronous, standard Fourier-based frequency estimation methods cannot be reliably applied. We propose here an alternative method using CWT.

The speech signal was first lowpass filtered to 3000 Hz using Butterworth 3rd order filter, energy envelope was then extracted by taking the absolute value of the signal, and the envelope was resampled to 200 Hz. Smoothing was further applied to remove the sub-syllable phonetic detail from the resulting signal using a Gaussian filter (the black curve at the bottom in Fig. 2). CWT energy (square of amplitude) scalogram was then calculated for the signal, with complex Paul mother wavelet, offering a suitable compromise between frequency and time resolution [17] (top panel in Fig. 2). Center of gravity of energy for each frame – approximating the most dominant rate component across all hierarchical levels – was used as a local speaking rate estimate (the black curve at the top panel in Fig. 2).

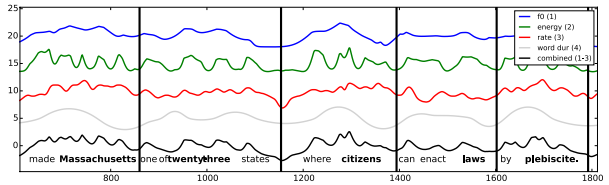

Figure 3: Prosodic feature signals used by the analyzed system, extracted for a sentence from BURNC. From the top, f_0, gate, wavelet-based rate signal and annotation based word duration signal. The black curve at the bottom shows a composite signal combining f_0, gain and rate features.

Fig. 3 shows examples of all four prosodic feature signals used in this work for a sentence from the BURNC corpus.

Finally, two composite parameter signals were created by combining f_0 and gain signals with, alternatively, the word du-

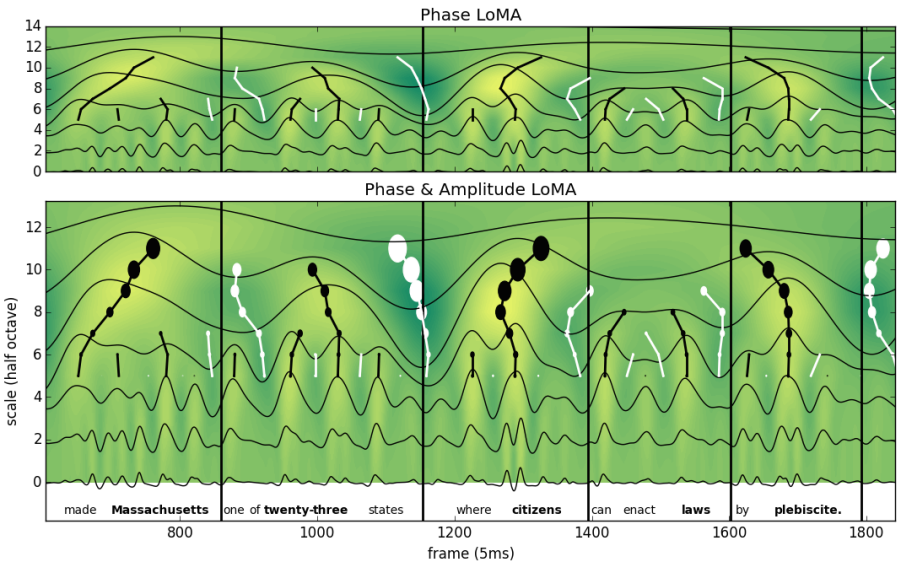

Figure 4: An example of lines of minimum (white) and maximum (black) amplitude extracted from scalogram of a composite acoutic feature signal. For comparison, the *phase* and *depth* methods are shown in the top and the bottom panel, respectively.

ration or speaking rate (the combined signal in Fig. 3 is an f_0–gain–rate composite). Each individual feature signal for an utterance was normalized to zero mean and unity variance, and the relevant three normalized signals were summed.

2.2. Phase and depth based boundary detection

CWT was performed on each composite signal using the second derivative of Gaussian (Mexican hat) mother wavelet, with a half-octave scale separation. Lines of minimum amplitude were then estimated from the scalogram, recursively connecting scale minima across multiple scales (see [18] for details)[1].

As seen in Fig. 4, the lines of minimal amplitude (shown in white) identify intervals where several "neighboring" scales reach minima approximately simultaneously. This means an inhibition of activity associated with the acoustic features under analysis (lowering f_0, gain and tempo) at multiple levels of prosodic hierarchy simultaneously. As such these events suggest a presence of a prosodic boundary.

Prosodic boundary is assumed to lie between speech intervals (phrases) associated with a greater activity pattern, i.e., one or more prominent speech event. These are indicated in Fig. 4 by lines of *maximum* amplitude highlighted in black. We have used this insight for boundary detection. First, we use purely the lines of minimal amplitude (i.e., connected phase resets in scale signals) that lie between two lines of maximal amplitude. The boundary strength is associated with the lengths of these lines of minimal amplitude. We refer to this techniques as *phase method* (top panel in Fig. 4).

In addition, the measure of boundary strength can be supplemented as a cumulative depth of the line of minimum amplitude. Using this additional information leads to a technique called here *depth method*.

[1]This process uses annotated word boundaries for identification of an appropriate world-level scale, but average of the unsupervised speech rate could be used without loss in accuracy.

3. Results

The boundary detection methods were evaluated on BURNC corpus (as this corpus was also used for the state of the art methods used for baseline comparison). Almost all of the annotated data were used for the experiment, totalling 442 stories or 29774 words. Word level break labels were derived by combining the time aligned syllable and word labels. Manually corrected alignments were used when available. The task was to predict the presence or absence of boundaries with break indices 3 or 4 between any two words. Continuous valued boundary predictions were converted to binary, by finding the best performing dividing point in terms of accuracy, using random 10 % of the data.

Tab. 1 lists the word-level results in terms of percentage of correct detections (accuracy) as well as precision, recall and *F*-score. As baselines, we report the majority class (predicting no boundary after each word) and the unsupervised and supervised techniques discussed in the Introduction. Using the same corpus as the baseline studies make our results reasonably comparable despite possible small differences in terms of the subset of the corpus used.

The results show that the presented CWT-based method compares favorably with the previous boundary detection techniques. The *depth* method using word duration information actually provides higher accuracy than supervised method of [3]. Moreover, the performance of *depth* methods using only acoustic information is comparable to that of the unsupervised technique of [4] that uses explicit syllable duration as well as lexical and syntactic information.

The *depth* method also performs considerably better than the *phase* one. This suggests an importance of quantitative information regarding the surrounding context: boundaries are not signaled merely by coordinating phase resets at multiple hierarchical levels by also by the way the resets are realized, the "depth" of the valley created by the parallel declination in multiple signal dimensions and hierarchical levels.

Although the CWT-based rate extraction does not yield de-

Table 1: Accuracy, F-value, precision and recall as evaluated for all boundary detection methods described here. Baselines: majority class and state-of-the-art supervised and unsupervised methods.

Method (*features*)	**Acc. %**	**F**	**Prec.**	**Rec.**
Phase ($f_0 g$)	77.1	0.56	0.61	0.51
Phase (f_0+*g*+*word*)	83.5	0.69	0.73	0.64
Phase (f_0+*g*+*rate*)	78.6	0.58	0.65	0.52
Depth (f_0+*g*)	81.3	0.57	0.80	0.45
Depth (f_0+*g*+*word*)	85.7	0.72	0.80	0.65
Depth (f_0+*g*+*rate*)	82.1	0.58	0.84	0.44
Baselines (*features*)				
Majority	72.0			
Sup'd (f_0+*g*+*word*) [3]	84.6		*	
Unsup'd (f_0+*g*+*syll*) [4]	81.1	0.64	0.69	0.66

*False positives rate of 9.11 % reported instead of F-value, precision and recall.

tection accuracy at the same level as explicit duration of individual words, it nevertheless provides some improvement over the methods using only f_0 and gain. It shows that although the rate extraction uses wavelet analysis, the technique subsequently used *again* for identification of lines of minimal amplitudes and boundary strength, the rate estimation as conceptualized here provides additional information to the system.

Examining the discrepancies between the manual annotations and the predictions based on wavelet analysis, two tendencies emerge. First, the boundaries with high tones are often not identified by the detection system, as the pitch movement goes against our simplified assumption that acoustic features are inhibited at phrase boundaries.

Second, there are many cases where the boundary is found, but not in the exact location identified by annotators. The acoustic boundaries tend to be fuzzy; instead of an exact boundary point, there appears to be a boundary region, sometimes spanning multiple syllables. In these cases, the continuous word duration feature works as an effective remedy as it encodes speaking rate changes in a way that highlights the (English) tendency of placing boundaries between long content words and short function words [19]. It is also possible that this tendency influenced the annotators in the corpus in ambiguous cases.

4. Discussion

The results here show that prosodic structure can – and probably should – be studied and represented in a unified framework comprising multiple relevant signal variables and multiple levels of speech hierarchy. In particular, they indicate that phrasal boundary detection (by automatic systems and, likely, also by the human listeners) is assisted by phenomena linked to hierarchical nature of speech as revealed by CWT analysis[2].

It is plausible that the boundary detection system based on the techniques described here could benefit from incorporating additional acoustic dimensions such as voice quality features; laryngealization, for example, is known to be associated with sentence and phrasal boundaries [20].

[2]Although not explicitly analyzed in this work, we found that the detection systems using the same features without CWT perform with approximately 4 % less accuracy.

The unsupervised wavelet-based rate estimation method, although not performing as well as the word duration signal, shows enough potential to warrant further analysis and development. For example, an adapted CWT technique could be used to extract multiple rate signals reflecting rate information at several levels of speech hierarchy simultaneously.

The boundary detection system can also profit from another acoustics-based speaking rate estimation methods [15, 16] or from using different signal representations instead of the low-pass filtered envelope used here.

Overall, our results indicate that utilizing purely acoustic features in an unsupervised way is a viable option for boundary detection. At the same time, they suggest that some degree of top down information (such as word duration) is probably necessary to reach detection precision achieved by supervised systems.

5. Acknowledgements

The work was partly funded by a Finnish Academy postdoctoral grant to the second author.

6. References

[1] "Simple4All speech synthesis project (EU-FP7)," 2014. [Online]. Available: http://www.simple4all.org

[2] A. Rosenberg, *Automatic detection and classification of prosodic events*. Columbia University, 2009.

[3] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 216–228, 2008.

[4] S. Ananthakrishnan and S. Narayanan, "Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling." in *INTERSPEECH*. Citeseer, 2006.

[5] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[6] I. Daubechies *et al.*, *Ten lectures on wavelets*. SIAM, 1992, vol. 61.

[7] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the teager energy operator," *Signal Processing Letters, IEEE*, vol. 8, no. 1, pp. 10–12, 2001.

[8] A. Alani and M. Deriche, "A novel approach to speech segmentation using the wavelet transform," in *Signal Processing and Its Applications, 1999. ISSPA'99. Proceedings of the Fifth International Symposium on*, vol. 1. IEEE, 1999, pp. 127–130.

[9] T. Irino and R. D. Patterson, "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-mellin transform," *Speech Communication*, vol. 36, no. 3, pp. 181–203, 2002.

[10] M. Vainio, A. Suni, D. Aalto *et al.*, "Continuous wavelet transform for analysis of speech prosody," *TRASP 2013-Tools and Resources for the Analysys of Speech Prosody, An Interspeech 2013 satellite event, August 30, 2013, Laboratoire Parole et Language, Aix-en-Provence, France, Proceedings*, 2013.

[11] A. S. Suni, D. Aalto, T. Raitio, P. Alku, and M. Vainio, "Wavelets for intonation modeling in HMM speech synthesis," in *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*, 2013.

[12] M. S. Ribeiro, J. Yamagishi, and R. A. Clark, "A perceptual investigation of wavelet-based decomposition of f0 for text-to-speech synthesis," in *Interspeech*, Dresden, Germany, 2015.

[13] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "Hmm-based speech synthesis utilizing glottal inverse filtering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 153–165, 2011.

[14] A. Suni, D. Aalto, and M. Vainio, "Hierarchical representation of prosody for statistical speech synthesis," *arXiv preprint arXiv:1510.01949*, 2015.

[15] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2190–2201, 2007.

[16] V. K. Prasad, T. Nagarajan, and H. A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Communication*, vol. 42, no. 3, pp. 429–446, 2004.

[17] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological Society*, vol. 79, pp. 61–78, 1998.

[18] M. Vainio, A. Suni, and D. Aalto, "Emphasis, word prominence, and continuous wavelet transform in the control of HMM-based synthesis," in *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*. Springer, 2015, pp. 173–188.

[19] M. Y. Liberman and K. W. Church, "Text analysis and word pronunciation in text-to-speech synthesis," *Advances in speech signal processing*, pp. 791–831, 1992.

[20] J. Kreiman, "Perception of sentence and paragraph boundaries in natural conversation," *Journal of Phonetics*, vol. 10, no. 2, pp. 163–175, 1982.