Coordination between $f_0$, intensity and breathing signals

Abstract. The present paper presents preliminary results on temporal coordination of breathing, intensity and fundamental frequency signals using continuous wavelet transform. We have found tendencies towards phase-locking at time scales corresponding to several prosodic units such as vowel-to-vowel intervals and prosodic words. The proposed method should be applicable to a wide range of problems in which the goal is finding a stable phase relationship in a pair of hierarchically organised signals.

1. Introduction

The question of the pulmonary contribution to speech prosody is by no means a new one. In his review of the field, Ohala (1990) traces its origins as far back as Scaliger (1610). But old questions have a curious property of being repeatedly reopened, inviting continual revisions. More recently, the issue was revisited by Lieberman (1967), who claimed that $f_0$ movement associated with prominence is primarily produced by momentary increase in exhalatory effort rather than by the action of laryngeal muscles. This hypothesis, however, was soon disproved. It has been shown that the effect of respiratory effort on $f_0$ is far too small to account for the variations observed in speech. More importantly, part of the variation of subglottal pressure is likely to be due to glottal influences, for instance changes in glottal area associated with $f_0$ modulation. At the same time, subglottal pressure may account for some of the variation in intensity, a finding consistent with the well-known relationship of expiratory effort and speech loudness.

Unlike the effect of expiratory effort on $f_0$ variation, the relationship between respiration and dynamic aspects of stress related to loudness have not been entirely settled. This idea is most commonly attributed to Stetson (1928), who proposed that syllables coincide with breath pulses, a momentary increase in expiratory effort\(^1\), which are amplified by the presence of stress. While subsequent studies have not verified the presence of breath pulses in stressed syllables, some authors have found evidence for respiratory markers of stress (e.g. Fonagy, 1958; Ladefoged, 1962; van Katwijk, 1974). As demonstrated by Ohala (1977), however, the interpretation of these findings is further complicated by possible supraglottal influences on subglottal pressure, e.g., due to increased airflow during production of fricatives or closure during plosives.

\(^1\)Stetson himself believed that any $f_0$ variation on stressed syllables is a by-product of changes in respiratory effort.
Nevertheless, Ohala did find some evidence for respiratory control of emphatic stress (but not for non-emphatic stress).

More recently, Slifka (2003) found that a peak in alveolar pressure coincides with the first stressed syllable in an utterance, but only if the stress falls early, that is on the first or the second syllable in that utterance. In addition, the pressure peak was found to be a robust temporal anchor point for other phonation parameters, such as phonation onset time. More recently, Isei-Jaakkola (2011) visually evaluated the alignment of maxima in the respiratory signal with peaks in the speech waveform envelopes in Chinese, Japanese and English, but her results were largely inconclusive. The debate is perhaps best summarised by Hayes’s (1995) comment that breath pulses are a sufficient but not a necessary condition for stress production.

The earlier studies have produced only weak evidence in favour of a respiratory contribution to stress production. However, it should be pointed out that without exception they were based on a limited number of speakers and have largely relied on visual inspection of respiratory traces alongside \( f_0 \) and amplitude trajectories. This method is likely to miss hierarchical dependencies between prosodic levels (segments, syllables, feet, etc.), as well as temporal dependencies when respiratory events precede or follow \( f_0 \) and/or intensity events. In this paper we revisit the problem using continuous wavelet transform (CWT). CWT is a particularly well suited for this task as it allows frequency decomposition of an input signal, effectively breaking up a time series into hierarchical components. Notably, when applied to the speech signal, CWT has been shown to identify levels corresponding to major prosodic constituents. We therefore use CWT to identify frequency prosodic levels where respiration, \( f_0 \) and intensity are most in-sync.

2. Material

The speech material for the present study consisted of 40 Swedish transitive verbs embedded in a carrier phrase "Mannen {verb} kvinnan" (The man {verb} the woman). Each sentence was produced three times with focal accent on the first, second and last word, elicited using appropriate questions. In the present study only sentences with focal accent on the verb were analysed.

The participants were two native speakers of Central Swedish. Their breathing was captured with Respiratory Inductance Plethysmography (RIP), which measures changes of cross-sectional area of the chest and the abdomen by means of two elastic belts. Prior to the recording, relative gains of the individual belts were calibrated using the isovolume manoeuvre (Konno & Mead 1967).
Subsequently, the signals from the two belts were added and the resulting signal, proportional to total lung volume change, was used in the analyses. The audio signal was recorded using a close-talking directional microphone (Sennheiser HSP 4). Data collection took place in a sound-treated studio in Phonetics Laboratory, Stockholm University. $f_0$ and intensity were extracted from the audio signal using the GlottHMM framework and post-processed using a gap-filling method (Suni, Šimko, Aalto, & Vainio, in press).

3. Methodology

3.1. Continuous Wavelet Analysis

Continuous Wavelet Transform (CWT) is a time-scale representation of a signal that presents a natural way of capturing the hierarchical prosodic structure of speech. This technique emerged independently in physics, mathematics, and engineering, and is currently widely used for analysis of complex signals including electro-physiological, visual and acoustic signals (Daubechies, 1992), including several areas related to speech prosody (Kruschke & Lenz, 2003; van Santen, Mishra, & Klabbers, 2004). CWT performs a decomposition of the signal by convolving it with a series of scaled versions of the mother wavelet function. The mother wavelet can be pictured as a “brief oscillation” with a short span of non-negligible activity referred to as the wavelet’s pseudoperiod. Convolution of the signal with the wavelet in essence captures the degree of the local similarity between the signal and the wavelet and thus highlights the events with duration corresponding to the wavelet’s pseudoperiod. Scaling of the wavelet, i.e., adjusting its pseudoperiod, allows for time-frequency representation of the signal at a wide range of (pseudo)frequencies. Unlike in the case of FFT, the window size used for this frequency analysis is “automatically” (and, in some sense, optimally) adjusted for different frequencies; this facilitates signal decomposition into components corresponding, in principle, to an arbitrarily broad band of frequencies from fast oscillations (e.g., those capturing $f_0$ frequency) to slow signal modulation reflecting prosodic structure of speech (syllables, prosodic words, phrases…).

Figure 1 shows an example of CWT analysis of the breathing signal and intensity contour for one of the utterances analysed in this paper. The top panels depict the waveform of the utterance with the analysed signals plotted below. The bottom panels show the results of CWT analysis using a complex Morlet mother wavelet scaled with pseudoperiods with the range between 125 ms and 1 s (cf. Torrence & Compo, 1998). The scalograms in the form of heat maps (with brighter areas corresponding to higher values, i.e. greater local similarity) are also presented in
the bottom panels. They are superimposed with scale functions, convolutions of the signal with wavelet scaled to three different pseudoperiods of approx. 200, 360 and 750 ms from top to bottom. The peaks of these functions roughly correspond to events manifested in the signals that are associated with hierarchical constituents of syllables, (prosodic) words and phrasal structure (focus in the present context).

In order to quantify the coordination level between $f_0$, intensity and breathing contours, the CWT decomposition was performed for all these signals for all the analysed utterances. Morlet mother wavelet (parameter $\sigma = 6$) was used with scales corresponding to pseudofrequencies spanning the interval 125 ms – 1 s. The scales were logarithmically (base 2) spaced with 32 scales per octave (97 scales in total).

**Figure 1.** CWT decomposition of breathing and intensity signals of the same utterance. The utterance waveforms (top), and breathing and intensity signals (middle). The scalograms in the bottom panels depict the results of CWT analysis, with the lighter areas corresponding to higher activity. Three scale functions (pseudoperiods of 200, 360 and 750 ms) are superimposed over the scalogram heatmaps.

3.2. Relative phase

One of the difficulties with evaluating links between respiration and surface characteristics of speech signals lies in the variability of temporal alignment of relevant events influenced by, for example, articulatory properties of speech segments. Unvoiced plosives obstruct the vocal tract
and can be assumed to be accompanied by a plateau of breathing activity followed by a small and relatively fast decrease in lung volume. Changes in lung volume that might accompany an increase in intensity and $f_0$ associated with, for instance, focus marking, might not necessarily temporally coincide with the corresponding $f_0$ and intensity events. In other words, the existence of coordination between breathing and prosodic variables does not imply that the relevant events in the compared signals are in phase.

Two oscillatory processes (of approximately the same periodicity) that are mutually coordinated in time, however, will exhibit a stable relationship between their phases. The relative phase, i.e., the difference between phases of the two signals, while not necessarily equal to 0, remains relatively constant during the period of coordination.

**Figure 2.** Relative phase extraction procedure. Left top and bottom: Breathing and intensity signal CWT components (with 200 ms pseudoperiods from Figure 1; in black) and their instantaneous phase (in grey). Left middle: Relative phase between the signals. Right: Distribution of relative phases in sample points shown as an angle (grey) and the circular mean (black).

In this work, we are using this insight in order to quantitatively evaluate the degree of coordination between breathing, $f_0$ and intensity signals at multiple prosodically relevant temporal scales in parallel. For each signal, the instantaneous phase at each sampling point was obtained for each scale function (see Figure 2; the phase is implicitly contained in the complex-valued scale functions resulting from using the complex mother wavelet). Subsequently, for each pair of signals – *breathing*—*intensity*, *breathing*—$f_0$ and *intensity*—$f_0$ – the corresponding phase values are combined (subtracted) to obtain relative phase (middle left plot in Figure 2). Finally,
for every scale, the circular mean and circular variance was computed from the empirical distributions of relative phase values across all utterances for each of the two speakers.

4. Results

In Figure 3 we plot circular variance of the relative phase between each signal pair (breathing—$f_0$, breathing—intensity, $f_0$—intensity) for a given scale and for each speaker. The measure can be interpreted as expressing stability of mutual coordination between the signals for events with the duration of the corresponding pseudo-period, with lower values indicating tighter temporal coordination between the two processes. Consequently, valleys (local minima) are timescales where the compared signals achieve the most stable phase relationship.

Figure 3. Circular variance of relative phase of the CWT components of pairs of signal (y-axis) is plotted against pseudoperiods of the wavelet used to extract the components (x-axis). Some of the local minima corresponding to the periods of relatively strong coordination between phases are highlighted by grey vertical lines.

For both speakers, the breathing—intensity coordination curve has two local minima, around 325 and 610 ms (for Speaker 1) and around 190 and 440 ms (for Speaker 2). For comparison, in Table 1 we present mean durations of vowel-to-vowel intervals (VTV) and individual words for the two speakers. Notably, the valleys in the breathing—intensity curve correspond almost perfectly to the mean duration of a VTV and of the second (focussed) word for the second
speaker (184 and 461 ms, respectively). In Speaker 1, who spoke more slowly than Speaker 2, the minima occur for longer pseudoperiods but they do not match this speaker’s average durations as closely as for Speaker 2. The location of the first local minimum for Speaker 1 exceeds the mean VTV duration by roughly 80 ms, and the second minimum exceeds the mean duration of the focussed word by approximately 40 ms. Nevertheless, the values are within the order of magnitude of the period of the underlying syllabic and focal processes. The first of these minima (pseudoperiod of 324 ms) corresponds to the mean duration of focally stressed VTVs with long vowels for that speaker.

Interestingly, Speaker 1’s mean duration values follow much more closely the breathing—$f_0$ curve with the first valley at 250 ms and a valley/plateau around 545 ms. There is also a third, more pronounced minimum at 692 ms, whose origin is difficult to establish. An extra valley is also present in the plot for Speaker 2 but it falls in between the “syllabic” and the “focal” valleys at 339 ms.

Table 1. Mean durations (ms) of vowel-to-vowel intervals and individual words for the two speakers. Duration of words were approximated using VTV boundaries.

<table>
<thead>
<tr>
<th></th>
<th>Speaker 1</th>
<th>Speaker 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>VTV</td>
<td>242</td>
<td>184</td>
</tr>
<tr>
<td>VTV (focal, short)</td>
<td>290</td>
<td>229</td>
</tr>
<tr>
<td>VTV (focal, long)</td>
<td>324</td>
<td>261</td>
</tr>
<tr>
<td>Pre-focal word</td>
<td>419</td>
<td>369</td>
</tr>
<tr>
<td>Focal word</td>
<td>564</td>
<td>461</td>
</tr>
<tr>
<td>Post-focal word</td>
<td>472</td>
<td>276</td>
</tr>
</tbody>
</table>

Finally, coordination between $f_0$ and intensity shows the clearest patterns with clearly discernible minima in both speakers. In addition, the location of the valleys seems to follow rather closely the minima in the individual breathing—$f_0$ and breathing—intensity curves, particularly for Speaker 2. For Speaker 1 the mapping is only approximate; however, the presence of two separate minima around 346 and 459 ms indicates that the minimum around 250 ms mark in the curves involving breathing signal may indeed correspond to different underlying processes.

5. Discussion

We have attempted to apply continuous wavelet transform to shed new light on an old problem
of the pulmonary contribution to speech production. We proposed that CWT with its hierarchical decomposition of the input signal is particularly suitable for tackling this problem. In addition, we have proposed a way of quantifying the degree of phase coupling between two signals by calculating circular variance of relative phase of these signals. Subsequently, we have applied this method to narrow-focus utterances produced by two Swedish native speakers. A comparison of their respiratory activity with $f_0$ and intensity traces revealed certain tendencies towards phase locking at timescales corresponding to several phonetic and linguistic constituents.

The primary contributions of this work are thus twofold. Firstly, we have put forward a method which should be applicable to a wide range of problems in which the goal is finding stable phase relationship in a pair of hierarchically organised signals. Secondly, we have used this method to provide preliminary results on pulmonary contribution to speech production.

Arguably, the most serious shortcoming of the current work is lack of proper statistical apparatus to establish robustness of the results. However, to the best of our knowledge none of the standard inferential methods (such as comparing the empirical distribution against a random baseline random using the Kolmogorov-Smirnov test) is immediately applicable to our material. Given the lack of a suitable inferential treatment, the present study can be regarded as merely a description (and one based on a very limited material at that) drawing parallels between mean durations of relevant phonetic constituents and tendencies of signal pairs to remain phase locked at a particular timescale. Admittedly, the choice of these constituents was somewhat arbitrary, but only partly so given existing evidence (or lack thereof) for the relationship between respiration and various prosodic constituents in earlier studies.

Specifically, we have found some tendencies for breathing, $f_0$ and intensity to be phase-locked at time scales roughly corresponding to syllable-sized chunks (VTVs), stressed syllables and words under narrow focus. Notably, the speakers and signal pairings did not show the same regularities, suggesting both interpersonal and systemic differences. For instance, some speakers might tend to display tighter coupling between respiration and $f_0$ or between respiration and intensity. In addition, certain aspects of speech production (lexical stress, narrow focus, etc.) may involve coordination of the three underlying processes to different degrees. Whether and which of these tendencies are present and statistically robust is a matter of future research.

At the same time, the uncovered coordination patterns tend to reflect differences in speech rate between the two speakers, indicating that they correspond to actual speech production patterns. The results are further corroborated by the fact that similar patterns emerge in the relatively uncontroversial pairing of $f_0$ and intensity. Finally, the fact that tendencies towards stable
configurations of relative phase have been found at timescales exceeding the duration of individual segments indicates that they are unlikely to be due to variations in airflow inherent in production of specific segments. In addition, we have found some indication not only of pulmonary action in producing focal accents but also in regular (non-focal) syllables and words. Finally, regardless of the limitations of the present approach, we consider it to be a significant improvement over methods used previously, which have largely involved visual inspection of two signals in search of gross similarities between them. We hope to have demonstrated that CWT allows for much finer analysis, both in temporal and frequency domains. That these come at a price of increased complexity is but a sad fact of life.

6. Acknowledgments

The research presented here was funded in part by a collaboration grant between Stockholm University and University of Helsinki, and in part by the Swedish Research Council grant 2014-1072 Andning i samtal (Breathing in conversation) to the second author.

7. References

Lieberman, P. (1967). Intonation, Perception and Language. MIT, Boston, MA, USA.
Academic Publishers.


