Intonation modelling using a muscle model and perceptually weighted matching pursuit

Pierre-Edouard Honnet[a], Branislav Gerazov[b], Aleksandar Gjoreski[b], Philip N. Garner[a]

[a] *Idiap Research Institute, Centre du Parc, Rue Marconi 19, PO Box 592, 1920 Martigny, Switzerland*
[b] *Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University – Skopje (FEEIT), Skopje, Macedonia*

Abstract

We propose a physiologically based intonation model using perceptual relevance. Motivated by speech synthesis from a speech-to-speech translation (S2ST) point of view, we aim at a language independent way of modelling intonation. The model presented in this paper can be seen as a generalisation of the command response (CR) model, albeit with the same modelling power. It is an additive model which decomposes intonation contours into a sum of critically damped system impulse responses. To decompose the intonation contour, we use a weighted correlation based atom decomposition algorithm (WCAD) built around a matching pursuit framework. The algorithm allows for an arbitrary precision to be reached using an iterative procedure that adds more elementary atoms to the model. Experiments are presented demonstrating that this generalised CR (GCR) model is able to model intonation as would be expected. Experiments also show that the model produces a similar number of parameters or elements as the CR model. We conclude that the GCR model is appropriate as an engineering solution for modelling prosody, and hope that it is a contribution to a deeper scientific understanding of the neurobiological process of intonation.

Keywords: Intonation modelling, matching pursuit, physiology, weighted correlation, text-to-speech synthesis

1. Introduction

We are interested generally in speech to speech translation (S2ST). At the time of writing, S2ST is becoming a reality; with both research (e.g., the U-STAR consortium[1]) and commercial (e.g., Skype[2]) systems being available. This is a consequence of the component technologies — automatic speech recognition (ASR), machine translation (MT) and text to speech synthesis (TTS) — becoming quite mature.

In the context of ASR, especially when the goal is to produce text, prosody is normally ignored. By contrast, in the context of TTS, production of appropriate prosody is necessary to approach the naturalness of human speech. Although some applications using TTS do not necessarily require a human sounding voice, many of them would be more attractive if the machine — or communication intermediary — was able to produce natural sounding speech.

In the case of S2ST, not only is a natural voice required, but also one that conveys the intent and nuances of the speaker. This includes the ability to correctly emphasise the words, or groups of words, according to what has been said in the source language. Of course, this places requirements on the MT component, be it a simple mapping or something more complex (Do et al., 2015; Anumanchipalli et al., 2012).

In the present study, we focus on intonation modelling. Intonation modelling can be seen as finding a "good" representation of the intonation signal. The challenges are then: *What should be this representation?* and, *How can its parameters be extracted?*

We recently proposed a model which can be both extracted from a speech signal and recreated in a synthetic speech signal (Honnet et al., 2015). The model is physiologically based and can be seen as a generalisation of the CR model, although differing in some aspects in its definition. Inspired by the work of Kameoka et al. (2010) on the prediction of the CR parameters, we define local components of intonation as impulse responses to critically damped systems. Our first approach consisted of extracting parameters with a standard matching pursuit algorithm, followed by a selection of extracted atoms based on their perceptual relevance. This work was concerned with minimising the reconstruction error and investigating different system orders for the model components.

In a second iteration (Gerazov et al., 2015), the perceptual relevance was integrated directly in the extraction process by modifying the cost function of the matching pursuit algorithm, yielding optimal local decomposition with respect to the perceptual measure used. The ability of the model to reach high perceptual similarity was investigated, and a comparison with the standard CR model was proposed, using a perceptually relevant objective measure.

Both approaches were validated on a rather small but multilingual dataset. In the present paper, we take the op-

[1] http://www.ustar-consortium.com
[2] https://www.skype.com/en/features/skype-translator/

portunity to consolidate our previous work, giving a more in-depth description of the model with a discussion on its physiological credibility; a detailed procedure for extracting parameters and a comparison with the standard CR model are also presented. Additionally, we present a more thorough evaluation, done on a much larger scale with a variety of speakers and languages. Some differences in the cost function for extraction are also introduced in Section 4.4.

In the following sections, after a review of background material, we expand the motivation for our model in terms of muscle modelling and put it in the context of the CR model. We go on to describe how the extraction can be done automatically, and in terms of perceptual metrics known to the linguistics community. Experiments are presented that evaluate the plausibility of the model and place it in the context of the state of the art.

2. Background

The need for correct intonation in TTS systems as well as the more general study of intonation have motivated the creation of different intonation and / or prosody models. In the context of TTS, adaptive systems — almost exclusively statistical parametric speech synthesis (SPSS) — are of great interest in the research community. The current state of the art systems for SPSS are based on hidden Markov models (HMMs) of Tokuda et al. (2002b) and Zen et al. (2009). HMM-based speech synthesis deals with intonation in a framewise manner; each frame from the training speech database has a value — or a null value in the case of an unvoiced frame — and HMM states are trained using these values. At synthesis time, F_0 is generated frame by frame, based on the HMM parameters.

Decision trees allow clustering of different features using different tree structure, thus one can expect that when clustering contextual features with respect to F_0, suprasegmental information in the label will have more impact than segmental information. However, this results in a speech often qualified as "flat" or lacking expressivity, which is due to the oversmoothing of HMMs (Toda and Tokuda, 2005).

There are three main ways of tackling the flatness of HMM-based synthesis at the intonation level: *i*) use a different representation of F_0 in the HMMs, *ii*) postprocess the synthetic intonation coming from HMMs, or *iii*) use an external prosody model that combines with other HMM parameters.

In the early stages of HMM-based synthesis, a multi-space probability distribution (MSD)-HMM was developed by Tokuda et al. (2002a) and became a standard way of handling the fact that speech can be voiced or unvoiced. More recently, some work was done using continuous F_0 and it was shown that continuous F_0 improves the perceived naturalness of synthesis (Yu and Young, 2011; Latorre et al., 2011). This was further improved by hierarchical modelling using a continuous wavelet decomposition to separate the different levels of variation in F_0 (Suni et al., 2013). In this work, the authors exploit the multi stream architecture of an HMM-based TTS framework to cluster these different temporal scale components with different decision trees.

In the second category, an example of what can be done to improve the output of HMM synthesis is given by Hirose et al. (2011, 2012). Based on the command response (CR) model of Fujisaki and Nagashima (1969), the idea is to estimate the F_0 model commands from linguistic information, and then optimise them according to the F_0 generated by HMMs. By modifying the estimated parameters, it becomes possible to increase the expressivity of the synthetic speech. Another attempt to integrate the CR model in HMM-based TTS was made by Hashimoto et al. (2012), where parameterised F_0, in respect to the CR model, was used for training the HMM intonation features. This improved the quality of the synthetic speech as the model smoothed the F_0 contour before training.

The external prosody models, or intonation models are numerous. They can roughly be divided into models that: *i*) model the surface pitch contour, and *ii*) integrate the underlying physiological mechanisms of pitch production. Most intonation models fall into the first group. The Tone and Break Indices (ToBI) model (Silverman et al., 1992) is not a true surface model, nor is it a physiological one. It is linguistically focused, but is underdetermined and contains annotation and pitch synthesis ambiguities. On the other hand, the Tilt model (Taylor, 2000) is specially tailored for automatic parameter extraction and pitch synthesis. It describes the pitch contour as a sequence of events with specific shapes that can be automatically extracted with an obvious resynthesis step. The INSINT (INternational Transcription System for INTonation) model (Hirst et al., 2000) expands on ToBI and allows for automatic parameter extraction. It models the MOMEL (MOdélisation de MELodie) stylised (Hirst and Espesser, 1993) intonation contour as a sequence of specific F_0 target points. The General Superpositional Model of Intonation (Van Santen and Möbius, 2000), models the pitch contour decomposing it into a sum of a microprosodic segmental perturbation, an accent and a phrase curve. Finally, the Superposition of Functional Contours (SFC) model (Bailly and Holm, 2005), is a data driven approach based on the superposition of intonation prototypes that are directly linked to linguistic information through the use of neural networks.

Only a few models actually try to explain the intonation by investigating its production aspect. The most popular model in this category is the command response (CR) model of Fujisaki and Nagashima (1969). This model decomposes the intonation into additive physiologically meaningful components. The CR model is attractive for two reasons:

1. it has a physiological explanation which tries to account for the underlying mechanisms behind intonation production, and

2. it has a mathematical form, which makes it possible to parameterise.

Extracting the model parameters from an F_0 contour is not trivial, but the opposite resynthesis operation is straightforward.

The qTA (quantitative Target Approximation) model (Prom-on et al., 2009), expands on the CR model, and uses pitch targets as input to the physiological model of pitch production. The StemML (Kochanski et al., 2003), on the other hand, imposes physiological constraints of smoothness and communication constraints specified by target accent templates to the modelling process.

3. Physiologically based intonation modelling

3.1. Motivation

In mimicking the abilities of humans in a machine, it is natural to try to mimic human physiological processes. It is certainly not necessary; this is evidenced by the fact that there are many speech recognition and synthesis methods that use physiologically implausible mechanisms (such as Markov models and windowed frames). However, doing so has two attractive possibilities: The first is the main goal of technological advancement; the second is one of scientific understanding of the underlying processes.

Further, it is clear that there are no fundamental differences between speakers of different languages. We may hence reasonably expect a physiological model not to be language dependent.

3.2. Sources of physiological variation in F_0

A detailed analysis of intonation production is given by Strik (1994). In this work, using electromyographic (EMG) recordings of the relevant laryngeal muscles, four physiological sources of F_0 change were identified by assessing their influence on pitch:

1. The *cricothyroid (CT) muscle* rotates the thyroid cartilage with respect to the cricoid, stretching the vocal folds and raising F_0.
2. The *vocalis (VOC) muscle* is found within the vocal folds; its contraction decreases vocal cord length, but increases their tensile stress, the net effect being a rise in F_0 (Titze and Martin, 1998).
3. The *sternohyoid (SH) muscle* is one of three strap muscles used to alter the position of the larynx; it lowers the larynx decreasing vocal cord tension and F_0.
4. The *subglottal pressure (P_{sb})* is found to linearly correlate with increased F_0.

The measurements presented by Strik (1994) show that the CT and VOC activations are correlated and cause a rise in F_0, as do peaks in P_{sb}. By contrast, the activation of SH coincides with drops in F_0. Another important observation to point out is that only the P_{sb} signal has a global component; the others feature only local ones.

3.3. The CR model in a muscle context

Fujisaki (2006) argues that there is a linear relationship between the value of the F_0 in the log domain and the contraction (length) of the laryngeal muscles discussed in Sec. 3.2; it follows that different components of F_0 are additive in the log domain. This leads to the CR model which describes the F_0 contour as a superposition of multiple components in the log domain: *i)* a base component related to the size and density of the vocal folds, which is constant for a given speaker, speaking style and emotional state, *ii)* a time varying global phrase component, and *iii)* a time varying local accent component. The last two components are associated with the activation of two parts of the CT muscle, both effecting a rise in the F_0 through a slow translatory movement and a fast rotary movement of the thyroid cartilage, respectively. These components are modelled as 2^{nd} order critically damped system responses to two types of positive excitation commands: impulses for the phrase component, and step functions for the accent component.

The CR model also allows for negative phrase and accent commands. Negative phrase commands are used to model for the phrase final drops in F_0 (Hirose and Fujisaki, 1982; Fujisaki and Hirose, 1984). On the other hand, the use of negative accent commands is limited to modelling tonal languages such as Mandarin and Thai, as well as pitch-accent languages, such as Swedish and Bengali (Fujisaki, 2006; Fujisaki et al., 1998; Fujisaki, 2004; Fujisaki et al., 1993; Saha et al., 2011). Both of these negative components are attributed to the opposite rotary movement of the thyroid via the thyrohyoid (TH) muscle (Fujisaki, 2006).

The CR model is in accord with the work of Strik (1994), with two notable differences:

1. Strik argues that the phrase global component is pneumatic in nature, while phrase final F_0 drops are due to local activation of the laryngeal muscles. This differentiates the two, rendering the use of negative phrase commands inconsistent with physiology.
2. The activations of the SH muscle were observed in Dutch in Strik's work, which is neither a pitch-accent nor a tonal language. This implies that negative accent components are more prevalent across languages and they should be integral to intonation modelling.

3.4. The HMM realisation of CR

In order to use the CR model as a way of synthesising intonation for TTS, a discrete-time version of the model was presented along with a statistical model for the F_0 contours by Kameoka et al. (2010). Hidden Markov models (HMMs) with a specific topology are used to model phrase and accent command generation with constraints: accent commands cannot overlap and phrase command cannot occur while an accent command is still active. The fact that this model uses HMMs makes it a good candidate to generate plausible intonation: even though it might not be the "right one", it should be "natural". The use of substate HMMs to model the duration of accent components was added later by Yoshizato et al. (2012).

3.5. Models of muscles

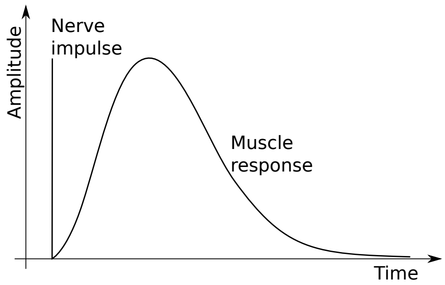

Figure 1: Hypothetical muscle twitch response to a nerve impulse. Loosely based on a diagram from Ruch et al. (1965).

It is known that signals are carried in nerves by means of impulses (or spikes) rather than by, say, absolute levels. When such spikes are applied to muscles, they result in a characteristic muscle twitch, illustrated in figure 1. This twitch can be thought of as the lowest level representation of muscular activity. Higher level movements, such as the contraction of a muscle, can be attributed to sequences of spikes with a period shorter than that of the twitch.

Perhaps the most obvious model of a muscular twitch is the impulse response of a system. In this case, the nerve spike can be thought of as the driving impulse and the system response is the response of the muscle. The simplest plausible case is the second order system arising from a spring-mass-damper arrangement. This is the same system model used in the CR model. In particular, critical damping leads to a gamma distribution shaped impulse response:

$$G_{k,\theta}(t) = \frac{1}{\theta^k \Gamma(k)} t^{k-1} e^{-t/\theta} \quad \text{for} \quad t \geq 0 \qquad (1)$$

The CR model uses such a damped system with order k assumed to be 2. However, Prom-on et al. (2009) showed that higher order models better model the vocal fold tension control. Other models, including the Hill model, are discussed by Gerazov and Garner (2015). Plamondon (1995) advocates the use of the log-normal distribution shaped response. This arises, via the central limit theorem, as a limiting case of many impulses travelling some distance from the brain to the muscle, and the muscle itself being compound.

Notice that, in a digital context, the response to a step function is equivalent to the impulse response to a train of impulses if the impulses are separated by exactly one frame. In this sense, the accent commands of the CR model can be viewed as sequences of impulses. This is part of the intuition behind the model of Kameoka et al. (2010) described above.

Accent components are then simply modelled using the same type of damped system as for phrase component with step functions replaced by impulse sequences. The log F_0 contour is then modelled as the sum of a base component and critically damped systems of the gamma form.

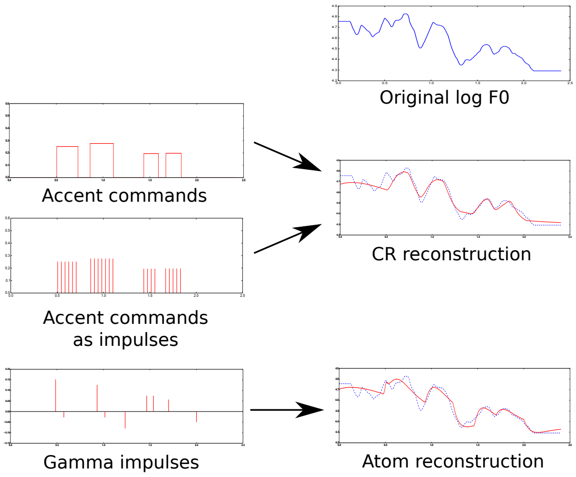

Figure 2: Illustration of different accent commands. Step functions can be represented as impulses; both lead to identical smoothed reconstructions. Removing the sequence constraint leads to a different smoothed reconstruction.

3.6. The atom-based intonation model

We have recently introduced an intonation model that is explicitly based on physiologically plausible muscle responses (Honnet et al., 2015; Gerazov et al., 2015). The model builds on Fujisaki's linear relationship between muscle contraction and log F_0 in assuming that the intonation contour can be represented as the superposition of gamma shaped *atoms*, being the muscular responses to individual nerve firings. Figure 2 sums up the reasoning behind our approach. In this figure, only local commands are displayed. The model is similar in spirit to that of Kameoka et al. (2010), but without the constraint that impulses must occur sequentially to form steps. Being free of such constraints, we refer to it as a *generalised* CR (GCR) model.

One compelling advantage of the GCR approach is that the extraction of the constituent atoms can be achieved simply via matching pursuit. The matching pursuit algorithm of Mallat and Zhang (1993) allows approximation of a signal as a linear combination of kernel functions — or atoms — taken from a dictionary. In an iterative manner, the algorithm finds the atom with the best correlation with the signal and then subtracts it until some desired accuracy is reached. This process reduces the reconstruction error by local optimisations.

Note that we take "physiologically plausible" to describe any of the models in the previous section, but without attempting to claim an exact physiological representation. Rather, the model is designed to be close enough to the physiology to have similar representational capability within the limits of Fujisaki's linear approximation and the matching pursuit algorithm.

3.7. Representational capability

If we use a critically damped second order system as the GCR kernel, it leads to a model with the same representational capability as the CR model. The proof of the equivalence follows from limiting cases of the driving functions: Consider just the accent commands. The CR model uses step functions; however, under PCM sampling, a step function is indistinguishable from a sequence of identical impulses spaced by the sampling period. So, the present model is able to represent anything that can be modelled by the CR model; this is similar to the motivation of Kameoka et al. (2010). Reciprocally, the step functions of the CR model can be narrowed such that they are the width of a single sample. In this case, a CR accent command is indistinguishable from the impulse of the present model, so the CR model can represent anything that can be modelled by the present model. Of course, in practice, these limiting cases do not occur; one goal of this paper is to investigate whether this is important.

3.8. Linguistic meaning

The GCR model is defined around a muscle model and an automatic extraction mechanism; it lacks a direct association with the underlying linguistic cues. This is in contrast to the CR model, which was designed specifically to model (Japanese) intonation patterns. Recently, however, this gap has been closed to some extent by studies by Delić et al. (2016), Szaszák et al. (2016) and Gerazov et al. (2016), who have shown that GCR atoms correlate rather well with ToBI markers, and can help with detection of emphasis (or stress). Honnet and Garner (2016) have also shown that GCR can be used to synthesise emphasis. Aside from being reassuring, this is intuitive in that ToBI is a mechanism for constructing linguistic cues; it is a semantic level below that of the cues themselves. It supports the GCR atoms being an appropriate input to a machine learning layer interfacing them to linguistic cues. This is beyond the scope of the present paper.

4. Perceptual matching pursuit

4.1. Introduction

At the outset, we used the default RMS error between atoms and the (logarithmic) F_0 contour as the metric for choosing atoms. However, we have little reason to believe that this is a perceptual measure. Not independent of this difficulty is the unnecessary modelling of unvoiced parts of F_0. Most intonation models use discontinuous pitch trackers and then interpolate unvoiced regions using, for instance, spline interpolation (Yu and Young, 2011). Then they concentrate the effort of modelling on the voiced parts (e.g., Mixdorff, 2000; Narusawa et al., 2002) or simply model everything equally (e.g., Hirst et al., 2000; Taylor, 2000). To avoid the latter, one needs a way of assessing which parts of the F_0 contour are perceptually relevant. Given a measure of perceptual relevance, where some segments are labelled as somehow more relevant than others, it is then possible to indicate that an algorithm should focus on the perceptually relevant segments whilst defocussing the less relevant ones. For instance, a-priori, we would expect the voiced segments to be more perceptually relevant.

4.2. Perceptually relevant objective measures of F_0 similarity

Two perceptually relevant objective measures of F_0 contour similarity — the weighted root-mean-square error (WRMSE), and the weighted correlation (WCORR) coefficient — were introduced by Hermes (1998). In this original formulation, the weighting function was defined as the maximum amplitude of the subharmonic sumspectrum (SHS), which is a weighted sum of the harmonics contributing to the pitch that was introduced by Hermes (1988).

The two proposed measures were aimed at automating the evaluation of student performance when teaching intonation (Hermes, 1998). The results showed that the measures correlated well with the similarity categorization done by five experienced phoneticians. Namely, the WRMSE was found to have a correlation of 0.679, to the experts' visual ratings, while the WCORR correlated better, at 0.67, with their auditory ratings. This is close to the interexpert agreement of 0.69 and 0.65 obtained for the two tasks. Moreover, approximate thresholds were calculated for classifying the perceptual similarity of two intonation contours using the objective measures. The thresholds for WCORR are given in Table 1. In our work we used modified versions of the WRMSE and WCORR to assess the perceptual similarity of our modelled pitch contour compared to the originally extracted F_0. The weighted RMS error (WRMSE) and the weighted correlation were calculated according to (2) and (3). Here f_0 is the reference F_0, $\hat{f}_0$ is the modelled F_0, i.e. its reconstruction, and $w(i)$ is the weighting function.

$$WRMSE = \sqrt{\frac{\sum_i w(i)(\hat{f}_0(i) - f_0(i))^2}{\sum_i w(i)}} \tag{2}$$

$$WCORR = \frac{\sum_i w(i)\hat{f}_0(i)f_0(i)}{\sqrt{\sum_i w(i)f_0(i)^2 \sum_i w(i)\hat{f}_0(i)^2}} \tag{3}$$

Table 1: Weighted correlation thresholds for perceptual similarity of two F_0 contours found by Hermes (1998).

Category	WCORR	Perceptual F_0 similarity
1	> 0.978	no differences
2	> 0.946	differences audible
3	> 0.896	differences clearly audible
4	> 0.827	linguistic differences
5	< 0.827	completely different

In our implementation we introduce three modifications to the original formulation.

1. We do not normalise the F_0 contours by their mean, as no offset is to be expected in our application scenario. The original implementation was formulated for a scenario where F_0 from different speakers were compared, requiring a normalisation for the register of each speaker. This is therefore not needed in our case.
2. We abandon the use of the equivalent rectangular bandwidth (ERB) scale of Glasberg and Moore (1990), in favour of using the logarithm of F_0; using the logarithm both has a long tradition in intonation modelling (Fujisaki and Nagashima, 1969), and is also equivalent to the semitones used in the perceptual intonation studies of d'Alessandro et al. (2011); Rilliard et al. (2011).
3. We define the weighting function to be (4), where $p(i)$ is the probability of voicing (POV), as defined by Ghahremani et al. (2014), and $e(i)$ is the energy contour of the speech signal. This is in accord with newer trends in perceptual intonation studies (Rilliard et al., 2011; d'Alessandro et al., 2011). It makes sense as regions of speech with higher energy and higher probability of voicing will have more impact on the perception of intonation — and speech, more generally. The introduction of a continuous POV estimate in (4), allows us to eliminate hard thresholds that were used to determine voicing (d'Alessandro et al., 2011) from our algorithm, making it more robust.

$$w(i) = p(i)e(i) \tag{4}$$

4.3. Atom selection using WRMSE

In our previous work, we used the WRMSE to give increased importance to the modelling of perceptually relevant segments of the F_0 contour (Honnet et al., 2015). To this end, we introduced an atom selection algorithm that uses the WRMSE to keep only the perceptually significant atoms from the set of atoms extracted using the approach outlined in Section 3. A summary of the procedure is given in Algorithm 1, where $F_{0\min}$ stands for the minimum value of F_0 for the given sentence, F_b is the base component, and F_{0p} is the phrase component.

The outlined algorithm proved to be adept at eliminating the extraneous atoms generated with the MP framework. This allowed for improved intonation modelling using the introduced gamma distribution-shaped atoms. The performance of the algorithm was verified across three different languages and six speakers (Honnet et al., 2015). Nonetheless, eliminating atoms at will from the set generated by the matching pursuit algorithm (MP) raised inconsistencies in the modelling process. Namely, sometimes when an atom which did not contribute significantly to the WRMSE was eliminated from the set, its influence in voiced regions was also eliminated. This means that the atoms that the MP algorithm fitted after it were then lacking in accuracy when modelling the F_0 contour. In other words, the subsequent atoms were compensating for, or taking into account, an atom that was not there anymore.

Algorithm 1 Atom decomposition with weighted RMSE based atom selection.

```
procedure ATOM DECOMPOSITION WITH WRMSE SELECTION
    Extract $F_0$, energy and $POV$ from waveform.
    Subtract $F_b = F_{0\min}$.
    Extract $F_{0p}$ using matching pursuit and subtract it.
    Extract atoms using matching pursuit.
Loop:
    if WRMSE $\leq$ Threshold then
        goto End.
    else
        if Atom decreases WRMSE by $> 0.001$ then
            Keep the atom and  goto Loop.
        else
            Discard the atom and goto Loop.
End.
```

4.4. Weighted correlation based atom decomposition

To improve our previous approach, we incorporated the perceptually relevant F_0 contour similarity measures, this time the weighted correlation defined in (3), as a cost function directly into the matching pursuit framework. The introduced weighted correlation atom decomposition (WCAD, Gerazov et al., 2015) algorithm directly extracts the atoms which are perceptually relevant, eliminating the need for subsequent atom selection. Another improvement in the algorithm is the introduction of a novel phrase atom extraction algorithm. These two key modifications make WCAD a more consistent, integrated algorithm, with added physiological plausibility.

4.5. Phrase atoms

The introduced phrase atoms are based on the qualitative shape of the global component of the subglottal pressure P_{sb} seen in the plots of the results obtained by Strik (1994). There, the global component starts with a peak at the start of phonation and then steadily decreases towards 0 with a time constant relative to the length of the utterance. The rise-time is much shorter than the fall-time, reflecting the nature of the physiological production of the P_{sb}, in which an initial pressure build-up that precedes speech is followed by its timely release that sustains phonation. This complex behaviour is provided by the interplay of the diaphragm and the rib cage muscles.

The phrase atoms are a modified version of the local atoms defined in (1), in that they follow one time constant θ_r during their rise, and another θ_f during their fall (5). Looking at Strik's plots, one can observe that the rise-time of the P_{sb} is consistent to a certain extent across the different utterances (Strik, 1994). Since we lack objective measurements to properly model the rise time, but we still need the rising part when modelling consecutive utterances, we use a fixed θ_r to represent a fast rise time across the phrase atoms. The exact value chosen is somewhat arbitrary, but should be in a range that corresponds to the measured rises in subglottal pressure in preparation for phonation. On the other hand, θ_f is chosen to maximise the cost function in the matching pursuit framework, as is outlined in Sec. 4.6. In (5), t_{rm} refers to the time instant in which the rising portion of the atom reaches its maximum, calculated according to (6). In the descending portion, the phrase atom starts from this maximum value and decreases towards 0. In order to compensate for the difference between t_{rm} and the maximum time instant t_{fm} of the fall function defined in 7, the time index t' is introduced in (5), calculated using (8), and illustrated in Figure 3.

$$G_{k,\theta_r,\theta_f}(t) = \begin{cases} \frac{1}{\theta_r^k \Gamma(k)} t^{k-1} e^{-t/\theta_r} & \text{for } 0 \leq t \leq t_{rm} \\ \frac{1}{\theta_f^k \Gamma(k)} t'^{k-1} e^{-t'/\theta_f} & \text{for } t > t_{rm} \end{cases} \tag{5}$$

$$t_{rm} = (k-1)\theta_r \tag{6}$$

$$t_{fm} = (k-1)\theta_f \tag{7}$$

$$t' = t - (t_{rm} - t_{fm}) \tag{8}$$

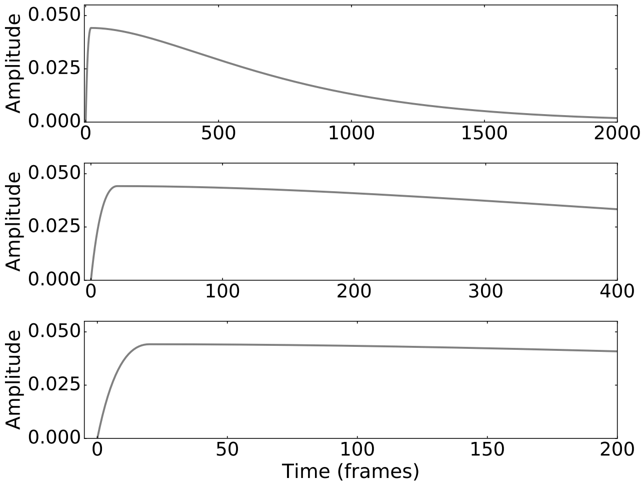

Figure 3: Modified phrase components. The rising part is much faster than the falling one. Different time scales are shown for the same phrase component to illustrate the effect at the sentence level (usually a few hundred frames).

4.6. WCAD algorithm

A summary of the Weighted Correlation Atom Decomposition (WCAD) algorithm is given in Algorithm 2. The algorithm integrates the weighted correlation in the calculation of the cost function of the matching pursuit algorithm, and accommodates the peculiarities of the phrase atom extraction.

At the start of the algorithm, the energy e and POV p are calculated from the waveform. These are then used to calculate the weighting function w using (4). Next, the phrase atom is extracted from the utterance. In concordance with Strik's findings, we fit a single phrase atom per breath group, i.e. one that fits the whole utterance, implicitly presuming that the utterance was spoken using a single breath. In the first step we estimate the start and end times of phonation, t_s and t_e, by thresholding the energy e with a starting threshold value T_s and a terminal threshold value T_e. The time instant t_s is used to align the position of the maximum of the phrase atom t_{rm} with the start of phonation in the utterance.

Algorithm 2 Weighted Correlation Atom Decomposition algorithm.

```
procedure WCORR ATOM DECOMPOSITION
    Extract $f_0, e$ and $p$ from waveform.
    Calculate $w$ from $e$ and $p$.
    Extract $t_s$ and $t_e$ of phonation.
    Find $\theta_f$ for phrase atom at position $t_s$ that maximizes WCORR · CORR for $t_s \leq t \leq t_e - t_{\text{off}}$.
    Calculate phrase atom amplitude using CORR.
    $f_{\text{diff}} = f_0 - \textit{phrase atom}$.
    $f_{\text{recon}} = \textit{phrase atom}$.
Loop:
    Find local atom with maximum WCORR · CORR with $f_{\text{diff}}$ for $t > t_s$.
    Calculate local atom amplitude using CORR.
    Increment atom count.
    $f_{\text{diff}} = f_{\text{diff}} - \textit{local atom}$.
    $f_{\text{recon}} = f_{\text{recon}} + \textit{local atom}$.
    if WCORR_norm of $f_{\text{recon}}$ > WCORR_norm thresh then
        goto End.
    else
        goto Loop.
End.
```

In the next step, θ_f is chosen to maximise the cost function calculated as the product of WCORR, as defined in (3), and the standard correlation function CORR, between the phrase atom and the F_0 contour. This is different to our previous work (Gerazov et al., 2015) where we used the WCORR itself as a cost function. This was introduced to circumvent deadlocks in the algorithm that would occur when the CORR function is zero in the location where the WCORR has a local maximum. In such cases, the algorithm would select the new atom based on the WCORR and give it zero amplitude based on the CORR. Thus, when it would get subtracted from the intonation contour, nothing would change, and the same atom would again be extracted based on the WCORR, repeating the process.

The cost function was calculated within the range of F_0 between t_s and $t_e - t_{\text{off}}$, where t_{off} is an offset time introduced to eliminate the phrase-final fall and rise in intonation from the phrase atom fitting. The extracted phrase atom amplitude is calculated using the standard correlation, after which the phrase atom is subtracted from f_0 to give the difference f_{diff}. The phrase atom is also used to initialise the F_0 reconstruction f_{recon}.

In the following part of the algorithm, local atoms are extracted from f_{diff} in a loop. At each iteration, the atom that maximizes the cost function WCORR · CORR the most is selected, disregarding the parts of f_{diff} before t_s and after t_e. Each atom is subtracted from f_{diff} before the next iteration, and also added to f_{recon}. The loop is repeated until either 1) the reconstruction WCORR reaches the selected threshold value, or 2) the chosen maximum number of atoms is reached.

We have chosen our stopping criteria to be the WCORR over the SNR (signal to noise ratio between the signal and the residual) used in MP, because of its determined perceptual significance, as was discussed in Sec. 4.2. Since the thresholds determined by Hermes (1998) are based on the original formulation in which the WCORR is calculated using the zero-mean versions of both f_0 and f_{recon}, we follow suit and substitute the normalised F_0 contours in (3), to obtain the WCORR$_{\text{norm}}$:

$$WCORR_{norm} = \frac{\sum_i w(i)(\hat{f}_0(i) - \hat{f}_{0m})(f_0(i) - f_{0m})}{\sqrt{\sum_i w(i)(\hat{f}_0(i) - \hat{f}_{0m})^2 \sum_i w(i)(f_0(i) - f_{0m})^2}}, \quad (9)$$

where $\hat{f}_{0m}$ and f_{0m} represent the respective means of the two contours. The WCORR$_{\text{norm}}$ is calculated for the part of the F_0 contour that was actually modelled by our WCAD algorithm, as bounded by t_s and t_e.

5. Evaluation

5.1. Experiment design

In conducting experiments, we aim to demonstrate two hypotheses. The first is about the model's ability to accurately capture the intonation dynamics, as well as the relative number of atoms required to reach a set modelling accuracy. Our hypothesis is that, because of the nature of the matching pursuit algorithm on which WCAD is built, our algorithm will progressively increase the WCORR with the addition of each of the atoms, reaching a saturation point at the optimal number of atoms. We also hypothesise that relatively few atoms will be needed to construct a model of the F_0 contour which is perceptually close to the actual F_0 contour with respect to our perceptually relevant weighted correlation coefficient.

The plausibility of the WCAD algorithm will be determined through assessing a) how well it can model the F_0 contour, and b) how many atoms does it need to do so. In order to determine this, we will analyse the contribution of each of the atoms as they are added in each iteration of the modelling procedure. More specifically, we will analyse how much does the addition of each atom increase the WCORR$_{\text{norm}}$ between the original and modelled F_0 contours. We will use the WCORR$_{\text{norm}}$, in order to assess the perceptual quality of the modelled F_0 using the thresholds discussed in Section 4.2. To extract the continuous F_0 and POV estimates we will use the pitch tracker implemented in Kaldi (Ghahremani et al., 2014)[3].

The second hypothesis is one of comparison of our generalised CR model with a state-of-the-art implementation of the standard CR model. We hypothesise that WCAD results would be comparable with those obtained with the CR model at a comparable number of atoms per syllable, and that the GCR model can ultimately reach higher accuracies than the CR model.

We will assess the comparative performance of our algorithm with the results obtained with the CR parameter extraction tool of Mixdorff (2000). We will calculate the WCORR$_{\text{norm}}$ obtained with the CR model, and use it to assess the perceptual quality of the modelled contour, comparing it with our WCAD results.

5.2. Data selection

The experiments were conducted on a large number of files, including speech in three different languages from both genders. This selection aims to demonstrate the language independent aspect of the model. Four databases were used: WSJ (Paul and Baker, 1992) and CMU Arctic (Kominek and Black, 2004) for English, BREF (Lamel et al., 1991) for French and Phondat (Hess et al., 1995) for German. This data can be seen as two main datasets:

- The CMU Arctic data consisted of two speakers: a male speaker, *bdl*, and a female speaker, *clb*. This set is aimed at evaluating the performance on the algorithm on the intra speaker variability aspect.

- The second set, using speech from many speakers and three languages, aims at evaluating the algorithm on multilingual and multispeaker aspects.

On the first dataset, from the utterances recorded for these 2 speakers, we manually selected the ones for which the used pitch extractor (Ghahremani et al., 2014) gave reliable results. The validity of the F_0 contours was assessed through comparison with two other pitch tracker outputs: STRAIGHT (Kawahara et al., 1999) and SSP from Garner et al. (2013)[4]. The final dataset totals 1729 utterances with a duration of 1.5 hours.

On the second dataset, a first random selection of the sentences was made, including 7085 sentences from WSJ, 15981 from BREF and 21587 from Phondat. To avoid using files for which the pitch tracker yields unreliable contours, we performed a pitch comparison using 3 different pitch trackers: SSP (Garner et al., 2013), the STRAIGHT vocoder (Kawahara et al., 1999) and the Kaldi pitch tracker

[3] See: http://kaldi.sourceforge.net/

[4] Available at: https://github.com/idiap/ssp

(Ghahremani et al., 2014). For all the files, the pitch was extracted with these 3 tools, and RMSE and correlation were calculated for each pair (Kaldi vs STRAIGHT, Kaldi vs SSP, SSP vs STRAIGHT). The files for which correlation was lower than 0.99 or RMSE was higher than 50Hz for at least one pair were discarded. As a result, 2453 files were selected for WSJ, 6387 for BREF, and 4433 for Phondat. Finally, to balance the subsets for each language, 8964 files were kept (2453 for WSJ, 2799 for BREF and 3712 for Phondat), by discarding the shortest (sometimes corresponding to single words) and longest files. The speech comes from 263 speakers: 76 for WSJ (37 males and 39 females), 23 for BREF (10 males and 13 females), and 164 for Phondat (78 males and 86 females). It amounts to a total of 12.6 hours (5.1, 4.9, 2.6). The final datasets are summarised in Table 2.

Table 2: Summary of test dataset sizes.

Group (lang.)	# speakers (M/F)	# sentences	Duration (h.)
Arctic (En)	2 (1/1)	1729	1.5
WSJ (En)	76 (37/39)	2453	5.1
BREF (Fr)	23 (10/13)	2799	4.9
Phondat (Ge)	164 (78/86)	3712	2.6
Total	265 (126/139)	10693	14.1

5.3. WCAD algorithm parameters

The parameters used in the WCAD algorithm were determined through qualitative assessment of its performance on a set of randomly chosen utterances from the first dataset (the CMU Arctic database (Kominek and Black, 2004)). It is reasonable to suppose that the optimal parameters are speaker dependent, but for the purpose of this paper we pool them together and assume speaker independence.

To determine the optimal order of our model k that is used in generating the gamma shaped phrase and accent atoms (1), the difference in WCAD performance for the various values of k was analysed. Fig. 4 shows the average WCORR versus the number of atoms per syllable for values of k in the range 2–7, for the French female speaker group. The curves were obtained by averaging the values over the whole French female speaker dataset. The curves are smooth because of the antialiasing of the plotting program, and the large amount of data. Fig. 8 shows the same measure for all the values for $k = 6$ with the mean curve. We can see that $k = 4, 5, 6, 7$ generally gives better performance than $k = 2, 3$. This is in the spirit of the findings of Prom-on et al. (2009), also discussed by Gerazov and Garner (2016). However, the high variance across the utterances makes it difficult to clearly favour one k. Moreover there is no plausible reason to use several orders in our model, so we assume that using order 6 is reasonable, as it gives a slightly better average performance than the k of 4 used in our previous work (Honnet et al., 2015), and the improvement when going to order 7 is small. For further discussion on the choice of order see the work of Prom-on et al. (2009).

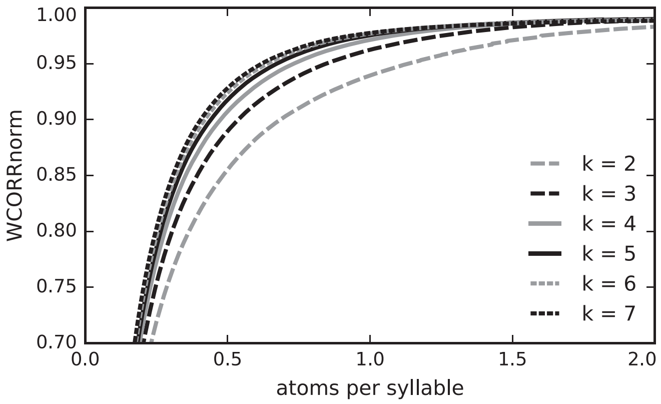

Figure 4: WCORR vs number of atoms per syllable for the French female speakers for different values of k.

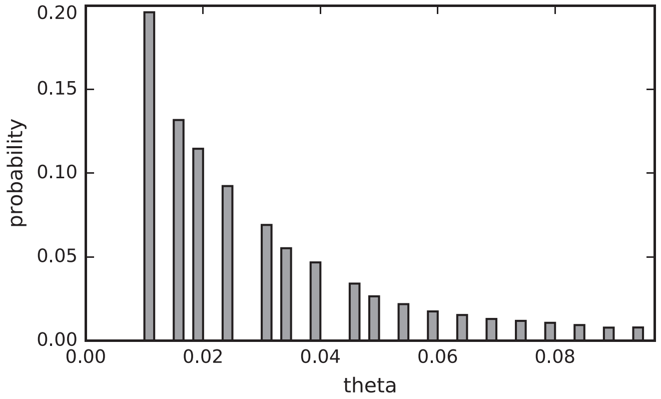

Figure 5: Histogram of the distribution of θ of the local atoms for the French female speakers.

To determine the start t_s and end t_e of phonation, we chose equal threshold values T_s and T_e of 0.01 for the normalised energy. The offset time t_{off} subtracted from t_e to leave out possible phrase-final falls and rises in F_0 was set to 150 ms. The θ_r for the rising part of the phrase atoms was fixed at 0.5. The range for the θ_f for the falling part of the phrase atoms was set to 0.1–10, and for the θ of the local atoms to 0.01–0.05. This way, the constructed dictionaries provide an atom variability sufficient for the function of the WCAD algorithm. The maximum θ_f of 10 covers the long utterances with a slowly decreasing global P_{sb} component. And the θ range encompasses the area where the values of θ concentrate, as can be seen in the histogram of their distribution in Fig. 5. The lower values of θ correspond to shorter atoms, which are mostly used for modelling sharper variations. The atoms using these low values have low amplitude; they help modelling the noise in intonation contours and getting higher accuracy in the reconstruction.

5.4. Example WCAD results

Example results of the Weighted Correlation based Atom Decomposition algorithm are given in Fig. 6 for the utterance *arctic_a0112.wav* taken from speaker *bdl*. The plots show the original F_0 contour, the extracted phrase atom and the extracted local atoms, and the reconstructed F_0. In order to obtain a clearer plot, only local atoms with amplitudes above 0.3 were used. As a comparison, the standard CR model extracted with Mixdorff's tool of the same example utterance is also given. We can see from the plots that the WCAD algorithm, with the limit put on the atom amplitude, extracts 1 phrase atom and 5 local atoms to model relatively well the F_0. Mixdorff's tool extracts 1 phrase command and 3 accent commands to model the same utterance. The lower number of components is advantageous, but the standard CR model, however, fails to capture the phrase-final drop in F_0. In fact, phrase-final drops were accounted for only later in the standard CR model, through the addition of negative phrase-final phrase commands (Fujisaki, 2004), and they are not automatically extracted by Mixdorff's tool.

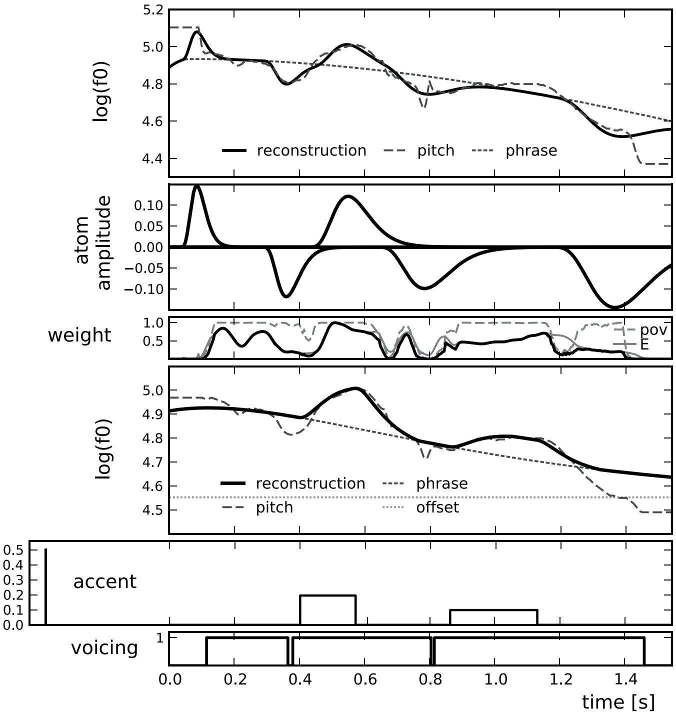

Figure 6: Reconstruction of F_0 contour using the Generalised CR model (1st panel), atoms extracted using the WCAD algorithm (2nd), and weighting function used (3rd); compared to the reconstruction using the standard CR model (4th), phrase and accent commands extracted using Mixdorff's tool (5th), and the voicing vector used (6th), for an utterance from *bdl*.

On the other hand, the lack of negative accent commands for English in the CR model, precludes the proper placement of the phrase component. An example of this can be seen in Fig. 7, in which the accent commands compensate for the wrongly placed phrase command. The WCAD algorithm, on the other hand, is not limited to using only positive local atoms, allowing it to do a better job at fitting the phrase atom, while at the same time being physiologically more plausible. For these two examples, we can see that the WCAD algorithm extracts a phrase component which looks like an average of F_0 movements. This smooth version of the F_0 curve makes sense physiologically as it would assure minimal activations, and thus conservation of energy. By contrast, the phrase component extracted by Mixdorff's algorithm is placed at the minimum of the F_0 curved. This can lead to incorrect accent commands (Fig. 7).

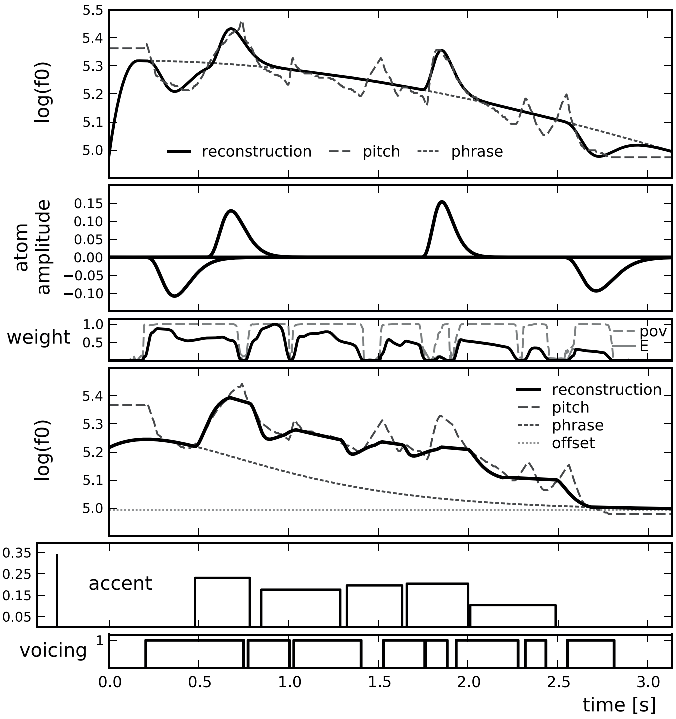

Figure 7: Reconstruction of F_0 contour using the Generalised CR model (1^{st} plot), atoms extracted using the WCAD algorithm (2^{nd}), and weighting function used (3^{rd}); compared to the reconstruction using the standard CR model (4^{th}), phrase and accent commands extracted using Mixdorff's tool (5^{th}), and the voicing vector used (6th), for an utterance from *clb*.

5.5. Results from experiments

In the examples shown in Figs. 6 and 7, we have limited WCAD to large amplitude atoms. The algorithm can, however, iteratively extract atoms to bring the modelled F_0 close to the original to an arbitrary degree, in terms of the cost function used. To analyse this performance we have calculated the $\text{WCORR}_{\text{norm}}$ at each iteration of the algorithm and plotted it as a point in the WCORR – atom/syllable plane, for all of the utterances for both speaker groups (male / female) from BREF. The results are shown in Fig. 8 as grey dots. The figure also shows the average $\text{WCORR}_{\text{norm}}$ relative to the number of atoms/syllable, averaged across all the sentences for each speaker group, as a black curve.

The average $\text{WCORR}_{\text{norm}}$ plots obtained for the different speaker groups from the multilingual set are plotted for comparison in Fig. 9. The curves represent the average performance of the GCR with $k = 6$ per speaker

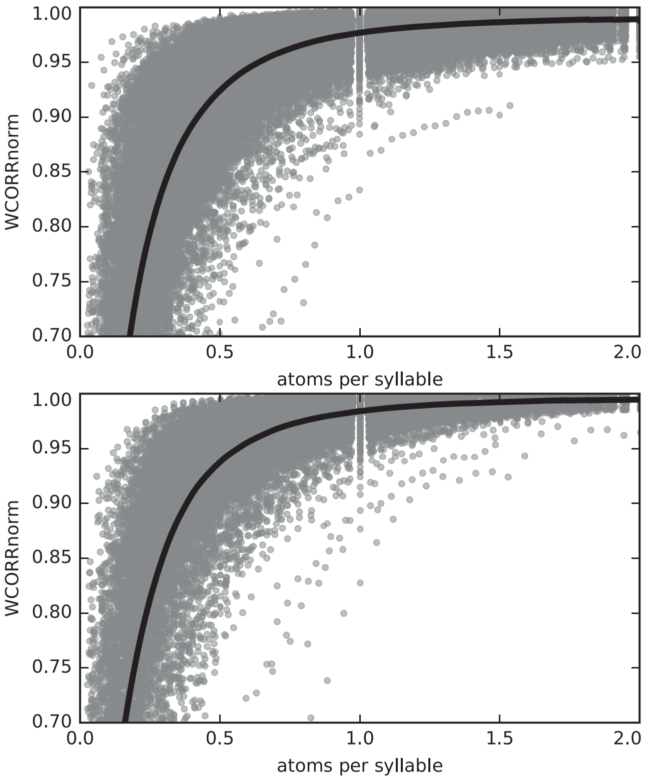

Figure 8: Weighted correlation of the zero-mean normalized F_0 contours relative to the number of atoms per syllable for all of the utterances for the French female (top) and male (bottom) speakers (gray points), and the calculated average curve (black), for $k = 6$.

group, while the dots represent the performance of Mixdorff's tool at the sentence level for the same speakers and sentences. In Mixdorff's case, each sentence is represented by a dot as it only gives one decomposition result. In the GCR case, according to the number of local component we extract, we get different WCORR, hence the average curves. To calculate the WCORR for the standard model we only used the part of the F_0 contour that was between the start and end of voicing.

We can see that, as hypothesised, at the start the WCAD algorithm gives rapid improvements in the $\text{WCORR}_{\text{norm}}$ with the inclusion of the first (larger) atoms in the model. The improvement in WCORR then gradually decreases as more (smaller) atoms are introduced. The plots show that the improvements in WCORR reach a saturation point around 1 atom/syllable for all of the speaker groups of all databases, hinting at a deeper link between the syllable unit and elementary intonation atoms. Such a hypothesis though, necessitates a thorough investigation that is beyond the scope of this paper.

The results show that the WCAD algorithm performs equally well for speakers of different languages and gender. The female speakers show a slightly lower performance of the GCR model, as they often have more variations in their intonation, requiring more components to get the same precision. The hypothesis that speaker and language play a role in the complexity of the patterns comes naturally,

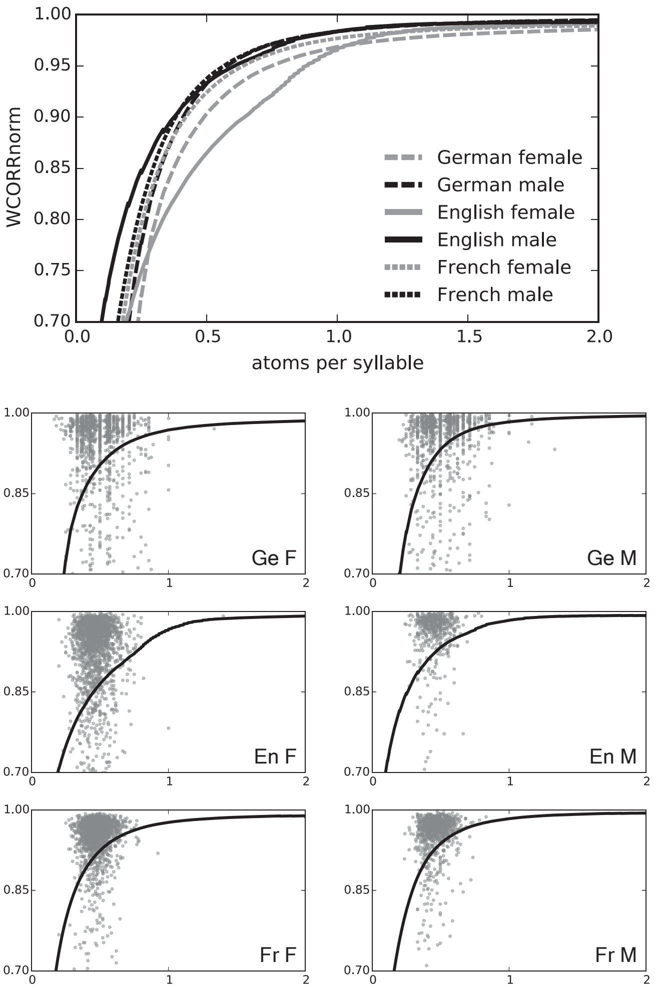

Figure 9: Average weighted correlation of the zero-mean normalized F_0 contours relative to the number of atoms per syllable for the different speaker categories from the multilingual set, for $k = 6$ (curves in the top panel). The WCORRs obtained with Mixdorff's implementation of the CR model are shown for comparison (dots) in the lower plots, for each speaker category.

however, the WCAD algorithm does not have an inconsistent behaviour across all of the data used, hinting at both its speaker and language independence.

Compared to the standard CR model, the WCAD algorithm underperforms when using a smaller number of atoms in some cases (first dataset), but its accuracy reaches and goes beyond that of the CR model as more atoms are added. It is important to note that in the case of the plotted points obtained from the CR model, "atoms/syllable" actually represents "commands/syllable", and that the commands in the CR model actually represent a response to a sequence of pulse excitations, as discussed in Sec. 3.2. On the other hand, the atoms in our generalised CR model correspond to single pulsed excitations, making straightforward comparison on this plot slightly biased.

In order to get a sense of the number of atoms/syllable needed for the WCAD algorithm to reach a certain perceptual accuracy in modelling the F_0 contour, we used the different WCORR perceptual thresholds presented in Table 1 as stopping criteria. The results of this analysis are given in Table 3. The table lists the average number of atoms/syllable needed to reach the different perceptual WCORR thresholds, for each of the speakers. We can see that to reach perceptual indistinguishability (Category 1) WCAD uses on average 1 atom per syllable for the first dataset, as was also hinted by the WCORR plots in Fig. 8. In the second dataset case, fewer atoms are needed to reach such perceptual quality. If we relax this accuracy condition and go with an F_0 model that allows for some perceptual difference (Category 2), the generalised CR model needs on average a bit more than half of this atom rate, i.e. 1 atom for every 2 syllables.

Table 3: Number of atoms/syllable needed on average to reach a chosen perceptual WCORR threshold, for each speaker group from all datasets.

Speaker group	Cat 1	Cat 2	Cat 3	Cat 4
bdl	0.75	0.48	0.34	0.24
clb	1.27	0.74	0.45	0.29
Average	1.01	0.61	0.39	0.26
En M	0.69	0.53	0.34	0.26
En F	0.90	0.62	0.47	0.34
Average	0.80	0.58	0.41	0.30
Fr M	0.71	0.47	0.32	0.22
Fr F	0.93	0.70	0.54	0.41
Average	0.82	0.59	0.43	0.32
Ge M	0.81	0.52	0.41	0.29
Ge F	0.79	0.60	0.41	0.33
Average	0.80	0.56	0.41	0.31

As a comparison to the performance obtained with the CR model, Table 4 gives the average WCORR, and the average total number of phrase and accent commands in the standard CR model for each speaker group. We can see that Mixdorff's tool on average gives a model with a WCORR of 0.96 on average for the first dataset, which corresponds to Category 2 from Table 1, and of 0.91 for the second dataset, which corresponds to Category 3. The average number of commands/syllable is 0.49 for the first dataset and 0.48 for the second. For a more readable comparison, Table 5 contains the number of components extracted with both models for the best category which could be reached using the CR model. Then, these average number of commands/syllable are to be compared with the results obtained with the WCAD algorithm at 0.61 for Category 2 (first dataset), and 0.43 atoms/syllable for Category 3 (second dataset). In the first dataset case, a few more atoms are required for the GCR model compared to the standard CR model for reaching the same perceptual quality, while for the second dataset which has more variability, the GCR requires fewer atoms than the standard CR model. This affirms the comparable performance of our algorithm.

Table 4: Average WCORR and number of commands/syllable obtained by the CR model, for each speaker group.

Speaker group	WCORR	Cat	commands	com/syl
bdl	0.96	2	5.7	0.48
clb	0.96	2	6.1	0.51
Average	0.96	2	5.9	0.49
En M	0.94	3	12	0.46
En F	0.91	3	14	0.47
Average	0.92	3	13	0.46
Fr M	0.95	2	12	0.46
Fr F	0.94	3	12	0.48
Average	0.94	3	12	0.47
Ge M	0.91	3	5	0.51
Ge F	0.83	4	5	0.50
Average	0.87	4	5	0.50

Table 5: Comparison of GCR and CR number of atoms/syllable for best category obtained with the CR model.

Speaker group	Category	com/syl CR	atom/syl GCR
En (Arctic)	2	0.49	0.61
En (WSJ)	3	0.46	0.41
Fr (BREF)	3	0.47	0.43
Ge (Phondat)	4	0.50	0.31

5.6. Discussion

The example figures demonstrate the qualitative advantages of the more flexible GCR model over the standard CR model. The allowance of negative atoms in the GCR model, as well as the design of the phrase atoms and the algorithm used to extract them, have allowed for the extraction of an observably better phrase component. These two advantages result in better, physiologically more plausible modelling results overall.

The experiments confirmed the plausibility of the GCR model, and the WCAD algorithm as a means for the extraction of its parameters. The results show that the model can successfully capture the intonation dynamics for different speakers and languages to an arbitrary precision. The built-in WCORR measurement allows the user to set the perceptual quality of the modelled intonation patterns. The results show that high perceptual quality can be obtained with the model when using around 1 atom per syllable.

The results from the comparison showed that the WCAD algorithm gives comparable modelling performance to the standard CR model with respect to our perceptually relevant measure at a given atom/syllable rate. It also accentuates the added flexibility of WCAD due to its iterative nature, which allows for an arbitrary modelling precision to be achieved. Namely, the results demonstrated that as more and more atoms are being added, the WCAD algorithm reaches WCORRs that can not reached by the standard CR model. This inherent flexibility in controlling the modelling accuracy of the GCR model allows users to tailor it to their modelling needs. For example, to infer basic linguistic meaning from the F_0 contour, one might opt for a smaller number of larger atoms in the decomposition. On the other hand, if an intonation generation algorithm should be trained using the GCR decomposition, then more "middle-sized" atoms should be added to capture all perceptually significant changes in pitch. Finally, if one is to alter the prosody of a speech signal, e.g., to synthesise emphasis (Honnet and Garner, 2016), then even small atoms should be extracted to retain verbatim pitch in the parts of the contour that are not to be altered.

An additional point that we need to emphasise when we compare the GCR to the CR is that the parameters of the GCR model can be extracted fully automatically using the proposed WCAD algorithm. On the other hand, there is no automatic way to extract the "right" parameters for the CR model. Even advanced tools such as Mixdorff's that we used, are prone to erroneous output and need expert adjustment.

However, achieving a high reconstruction accuracy with the WCAD algorithm introduces atoms which model the noise inherent to intonation curve. It is a difficult problem to automatically separate the prosodically meaningful events from the microprosody noise with no linguistic and paralinguistic information.

6. Conclusion

The GCR model is a physiologically plausible model of intonation. It can be trained in a completely automatic manner via matching pursuit. Further, the training process can be altered to depend on weighted correlation — a perceptually relevant measure — rather than simply RMS error. Weighted correlation also allows some choice of a perceptually relevant stopping criterion in the modelling process. Being based on the physiology of intonation production, the GCR model and the WCAD algorithm are inherently speaker and language independent.

Experimental results have shown that the model behaves broadly as expected in that it can model intonation dynamics well. Although arbitrary detail can be modelled by adding more atoms, an optimal number corresponds to the WCORR curve levelling out.

Experiments have further shown that the model compares well with the CR model in terms of number of parameters (or prosodic components) per syllable. This follows from the fact that both models have the same representational capability. In fact, the GCR model can outperform the CR model in terms of modelling accuracy, but may use more atoms in doing so.

Other researchers have shown that GCR atoms have linguistic meaning via their correlation to ToBI events, and can be used to help extract and synthesise emphasis. In the present work, we make no attempt to infer linguistic meaning; this is a topic for future research.

Although the model is physiologically *plausible*, the experiments certainly do not show that it is an *exact* model of the underlying physiological system; the shape, amplitude and frequency of atoms may all be different. This is also matter for future research. However, in addition to its merit as an engineering tool, we hope that the GCR model can serve as a framework for such research and deeper understanding of the neurobiological process of intonation formation as sum of elementary intonation atoms.

The software developed to implement the WCAD algorithm is freely available[5].

7. Acknowledgements

This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS) and under grant IZ73Z0 152495: the SCOPES Project on Speech Prosody (SP2).

Anumanchipalli, G. K., Oliveira, L. C., Black, A. W., 2012. Intent transfer in speech-to-speech machine translation. In: Proceedings of the fourth IEEE Workshop on Spoken Language Technology. pp. 153–158.

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

d'Alessandro, C., Rilliard, A., Le Beux, S., March 2011. Chironomic stylization of intonation. Journal of the Acoustical Society of America 129 (3), 1594–1604.

Delić, T., Gerazov, B., Popović, B., Sečujski, M., August 2016. A linguistic interpretation of the atom decomposition of fundamental frequency contour for American English. In: Ronzhin, A., Potapova, R., Németh, G. (Eds.), Speech and Computer. Vol. 9811 of Lecture Notes in Artificial Intelligence. Springer International Publishing, Budapest, Hungary, pp. 59–66, 18th International Conference, SPECOM 2016.

Do, Q. T., Takamichi, S., Sakti, S., Neubig, G., Toda, T., Nakamura, S., September 2015. Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs. In: Proceedings of Interspeech. Dresden, Germany.

Fujisaki, H., 2004. Information, prosody, and modeling-with emphasis on tonal features of speech. In: Speech Prosody 2004, International Conference.

Fujisaki, H., May 2006. The roles of physiology, physics and mathematics in modeling prosodic features of speech. In: Speech Prosody. Dresden, Germany.

Fujisaki, H., Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. Journal of the Acoustical Society of Japan (E) 5 (4), 233–242.

Fujisaki, H., Ljungqvist, M., Murata, H., 1993. Analysis and modeling of word accent and sentence intonation in Swedish. In: Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on. Vol. 2. IEEE, pp. 211–214.

Fujisaki, H., Nagashima, S., 1969. A model for the synthesis of pitch contours of connected speech. Tech. rep., Engineering Research Institute, University of Tokyo.

Fujisaki, H., Ohno, S., Wang, C., 1998. A command-response model for F0 contour generation in multilingual speech synthesis. In: The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis.

Garner, P. N., Cernak, M., Motlicek, P., January 2013. A simple continuous pitch estimation algorithm. IEEE Signal Processing Letters 20 (1), 102–105.

Gerazov, B., Garner, P. N., November 2015. An investigation of muscle models for physiologically based intonation modelling. In: Proceedings of the 23rd Telecommunications Forum. Belgrade, Serbia, pp. 468–471.

[5] https://github.com/dipteam/wcad

Gerazov, B., Garner, P. N., August 2016. An agonist-antagonist pitch production model. In: Ronzhin, A., Potapova, R., Németh, G. (Eds.), Speech and Computer. Vol. 9811 of Lecture Notes in Artificial Intelligence. Springer International Publishing, Budapest, Hungary, pp. 84–91, 18th International Conference, SPECOM 2016.

Gerazov, B., Gjoreski, A., Melov, A., Honnet, P.-E., Ivanovski, Z., Garner, P. N., September 2016. Unified prosody model based on atom decomposition for emphasis detection. In: Proceedings of ETAI. Struga, Macedonia.

Gerazov, B., Honnet, P.-E., Gjoreski, A., Garner, P. N., September 2015. Weighted correlation based atom decomposition intonation modelling. In: Proceedings of Interspeech. Dresden, Germany.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S., 2014. A pitch extraction algorithm tuned for automatic speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 2513–2517.

Glasberg, B. R., Moore, B. C. J., August 1990. Derivation of auditory filter shapes from notched noise data. Hearing Research 47, 103–108.

Hashimoto, H., Hirose, K., Minematsu, N., 2012. Improved automatic extraction of generation process model commands and its use for generating fundamental frequency contours for training HMM-based speech synthesis. In: Proceedings of Interspeech.

Hermes, D. J., 1988. Measurement of pitch by subharmonic summation. Journal of the Acoustical Society of America 83 (1), 257–264.

Hermes, D. J., February 1998. Measuring the perceptual similarity of pitch contours. Journal of Speech, Language, and Hearing Research 41 (1), 73–82.

Hess, W. J., Kohler, K. J., Tillmann, H.-G., 1995. The Phondat-verbmobil speech corpus. In: Proceedings of EUROSPEECH.

Hirose, K., Fujisaki, H., 1982. Analysis and synthesis of voice fundamental frequency contours of spoken sentences. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82. Vol. 7. IEEE, pp. 950–953.

Hirose, K., Hashimoto, H., Ikeshima, J., Minematsu, N., May 2012. Fundamental frequency contour reshaping in HMM-based speech synthesis and realization of prosodic focus using generation process model. In: Speech Prosody.

Hirose, K., Ochi, K., Mihara, R., Hashimoto, H., Saito, D., Minematsu, N., August 2011. Adaptation of prosody in speech synthesis by changing command values of the generation process model of fundamental frequency. In: Proceedings of Interspeech. Florence, pp. 2793–2796.

Hirst, D., Di Cristo, A., Espesser, R., 2000. Levels of representation and levels of analysis for the description of intonation systems. In: Prosody: Theory and experiment. Springer, pp. 51–87.

Hirst, D., Espesser, R., 1993. Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut Phontique d'Aix, 75–85.

Honnet, P.-E., Garner, P. N., September 2016. Emphasis recreation for TTS using intonation atoms. In: Proceedings of the 9th ISCA Speech Synthesis Workshop. Sunnyvale, CA, USA.

Honnet, P.-E., Gerazov, B., Garner, P. N., April 2015. Atom decomposition-based intonation modelling. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, Brisbane, Australia.

Kameoka, H., Le Roux, J., Ohishi, Y., September 2010. A statistical model of speech F0 contours. In: Proceedings ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA). pp. 43–48.

Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication 27 (3), 187–207.

Kochanski, G., Shih, C., Jing, H., 2003. Quantitative measurement of prosodic strength in Mandarin. Speech Communication 41 (4), 625–645.

Kominek, J., Black, A. W., 2004. The CMU Arctic speech databases. In: Fifth ISCA Workshop on Speech Synthesis.

Lamel, L. F., Gauvain, J.-L., Eskenazi, M., 1991. BREF, a large vocabulary spoken corpus for French. In: Proceedings of EUROSPEECH. pp. 505–508.

Latorre, J., Gales, M. J., Buchholz, S., Knill, K., Tamura, M., Ohtani, Y., Akamine, M., 2011. Continuous F0 in the source-excitation generation for HMM-based TTS: do we need voiced/unvoiced classification? In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4724–4727.

Mallat, S. G., Zhang, Z., 1993. Matching pursuits with time-frequency dictionaries. Signal Processing, IEEE Transactions on 41 (12), 3397–3415.

Mixdorff, H., 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vol. 3. Istanbul, Turkey, pp. 1281–1284.

Narusawa, S., Minematsu, N., Hirose, K., Fujisaki, H., 2002. A method for automatic extraction of model parameters from fundamental frequency contours of speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vol. 1. pp. 509–512.

Paul, D. B., Baker, J. M., 1992. The design for the wall street journal-based CSR corpus. In: Proceedings of the Workshop on Speech and Natural Language. Stroudsburg, PA, USA, pp. 357–362.

Plamondon, R., March 1995. A kinematic theory of rapid human movements: Part I: Movement representation and generation. Biological Cybernetics 72 (4), 295–307.

Prom-on, S., Xu, Y., Thipakorn, B., January 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. Journal of the Acoustical Society of America 125, 405–424.

Rilliard, A., Allauzen, A., Boula de Mareüil, P., August 2011. Using Dynamic Time Warping to compute prosodic similarity measures. In: Proceedings of Interspeech. Florence, Italy, pp. 2021–2024.

Ruch, T. C., Patton, H. D., Woodbury, J. W., Towe, A. L. (Eds.), 1965. Neurophysiology, 2nd Edition. No. ISBN 0-7216-7831-9. W. B. Saunders Company, West Washington Square, Philadelphia, Pa. 19105.

Saha, A., Basu, T., Warsi, A. H., Hirose, K., Fujisaki, H., 2011. Subjective evaluation of joint modeling of pause insertion and F0 contour generation in text-to-speech synthesis of Bangla. In: Oriental COCOSDA 2011. pp. 6–9.

Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B., Hirschberg, J., October 1992. Tobi: a standard for labeling english prosody. In: ICSLP. Vol. 2. Banff, pp. 867–870.

Strik, H., October 1994. Physiological control and behaviour of the voice source in the production of prosody. Ph.D. thesis, Dept. of Language and Speech, Univ. of Nijmegen, Nijmegen, Netherlands.

Suni, A., Aalto, D., Raitio, T., Alku, P., Vainio, M., August 2013. Wavelets for intonation modeling in HMM speech synthesis. In: 8th ISCA Workshop on Speech Synthesis. Barcelona, Spain, pp. 305–310.

Szaszák, G., Tündik, M. Á., Gerazov, B., Gjoreski, A., August 2016. Combining atom decomposition of the f0 track and HMM-based phonological phrase modelling for robust stress detection in speech. In: Ronzhin, A., Potapova, R., Németh, G. (Eds.), Speech and Computer. Vol. 9811 of Lecture Notes in Artificial Intelligence. Springer International Publishing, Budapest, Hungary, pp. 165–173, 18th International Conference, SPECOM 2016.

Taylor, P., March 2000. Analysis and synthesis of intonation using the tilt model. Journal of the Acoustical Society of America 107, 1697–1714.

Titze, I. R., Martin, D. W., 1998. Principles of voice production. Journal of the Acoustical Society of America 104 (3).

Toda, T., Tokuda, K., 2005. Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. In: Proceedings of Interspeech.

Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., 2002a. Multi-space probability distribution HMM. IEICE TRANSACTIONS on

Information and Systems 85 (3), 455–464.

Tokuda, K., Zen, H., Black, A. W., 2002b. An HMM-based speech synthesis system applied to english. In: Proc. of 2002 IEEE SSW. IEEE, pp. 227–230.

Van Santen, J. P., Möbius, B., 2000. A quantitative model of F0 generation and alignment. IntonationAnalysis, Modelling and Technology, 269–288.

Yoshizato, K., Kameoka, H., Saito, D., Sagayama, S., 2012. Statistical approach to Fujisaki-model parameter estimation from speech signals and its quantitative evaluation. In: Speech Prosody. pp. 175–178.

Yu, K., Young, S., July 2011. Continuous F0 modeling for HMM based statistical parametric speech synthesis. IEEE Transactions on Audio, Speech and Language Processing 19 (5), 1071–1079.

Zen, H., Tokuda, K., Black, A. W., 2009. Statistical parametric speech synthesis. Speech Communication 51 (11), 1039–1064.