ATOM DECOMPOSITION-BASED INTONATION MODELLING

Pierre-Edouard Honnet[1,2]*, Branislav Gerazov*[1]*, Philip N. Garner*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Switzerland

ABSTRACT

Current statistical parametric text-to-speech (TTS) synthesis methods allow production of neutral speech with acceptable quality. However, prosody is often qualified as unsatisfactory and sounding too flat. In this paper, we address intonation modelling for TTS based on physiological aspects of prosody production. A set of gamma distribution shaped atoms is defined and then intonation decomposition is performed using a matching pursuit algorithm. Some preliminary experiments show that this model allows easy extraction of physiologically meaningful atoms that could be used to generate intonation in a TTS system.

Index Terms— Intonation modelling, matching pursuit, physiology, text-to-speech synthesis

1. INTRODUCTION

We are interested in modelling prosody in the context of speech to speech translation. To this end, we require a model that can be both extracted from a speech signal and recreated in a synthetic speech signal. Further, given that prosodic events will need to be translated, it is necessary that the semantic events be clearly associated with their acoustic realisation.

Prosody modelling is a topic that has been investigated for many decades. With regard to TTS, it has become even more critical with the emergence of efficient statistical parametric TTS in the last years, as in recent systems, prosody is a limiting factor towards naturalness. Although neural networks recently started to be used for this purpose [1, 2, 3], hidden Markov model (HMM)-based TTS [4] remains the most popular way of achieving statistical parametric synthesis.

If we consider the three main aspects of prosody in a speech signal – intonation, duration and intensity – HMM-based TTS deals with intensity and intonation in a similar manner, as they are modelled framewise. Duration is modelled using hidden semi-Markov models (HSMMs) which encode duration as a parameter. The higher level dependencies of prosody are modelled using decision trees that take into account the context of each phone (position in syllable, in word, in phrase, stress, etc.). This framework allows synthesis of neutral read speech with satisfactory quality, but in the case of expressive speech, prosody often impacts the naturalness of the produced speech.

There are many intonation models, that we can divide in two categories: the ones which directly model fundamental frequency F_0 – e.g. [5, 6, 7, 8, 9] and the ones which try to imitate the underlying F_0 production process – e.g. [10, 11]. Our interest towards the latter category lead us to develop a model that is based on physiological evidence of prosody production. One of the motivations of such a model is that it would be language independent. The possiblity to extract information from intonation and to synthesise intonation from parameters also makes it appealing for speech to speech translation. We propose a model using a matching pursuit algorithm to decompose intonation into physiologically meaningful atoms and present some preliminary results.

The paper is organised as follows: Section 2 presents the underlying physiological aspect of prosody production, Section 3 introduces our model, Sections 4 and 5 give details on the experimental setup, evaluation and results, and Section 6 concludes the paper and gives perspectives.

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 SP2: the SCOPES Project on Speech Prosody.

2. PHYSIOLOGICAL ASPECT OF INTONATION GENERATION

The physiology of the production of intonation gives a solid basis for building an intonation model that will be language independent as the same vocal apparatus is used to generate pitch in all languages. The physiological mechanisms at work determining the frequency of vocal fold vibration are quite complex.

2.1. Sources of physiological variation in F_0

Four physiological sources of F_0 change were identified by Strik [12] where their influence on pitch was assessed through measurements including electromyographic (EMG) recordings of the relevant laryngeal muscles.

1. Cricothyroid (CT) muscle – rotates the thyroid cartilage in respect to the cricoid, stretching the vocal cords and raising F_0,
2. Vocalis (VOC) muscle – found within the vocal cords, its contraction decreases vocal cord length, but increases their tensile stress, the net effect being a rise in F_0 [13],
3. Sternohyoid (SH) muscle – one of three strap muscles used to alter the position of the larynx; lowers the larynx decreasing vocal cord tension and F_0,
4. Subglottal pressure (P_{sb}) – increased P_{sb} is found to linearly correlate with increased F_0.

The measurements presented by Strik [12] show that the CT and VOC activations are correlated and effectuate a rise in F_0, as do peaks in P_{sb}. In contrast, the activation of SH coincides with drops in F_0. Another important observation made by Strik is that only the P_{sb} signal has a global component, while the other feature only local ones. This leads him to argue that it is in fact the P_{sb} which is responsible for the phrase component of intonation.

2.2. Physiological interpretation of Fujisaki's model

Fujisaki's model [10] is based on modelling intonation using a global phrase component and local accent components. Seeking physiological interpretation of these two parameters, Fujisaki stated that they

are both attributed to the CT muscle, i.e. to its two functionally different parts [14]:

- *pars obliqua* – causes horizontal translation of the thyroid, stretches the vocal cords and raises F_0, responsible for the phrase component,
- *pars recta* – causes thyroid rotation as described in 2.1, responsible for the accent component.

The physiological basis for negative accent components corresponding to drops in F_0, which are needed to model tonal languages with Fujisaki's model, are attributed to the thyrohyoid (TH) muscle which, like SH, is one of the strap muscles and rotates the thyroid in the opposite direction to the one caused by the CT.

The physiological interpretation of Fujisaki's model is not in complete accord with the findings presented by Strik (section 2.1). One noteworthy difference is that Strik [12] has shown that negative local components are not exclusive to tonal languages, as they are clearly found in Dutch. This leads us to believe that a more physiologically plausible intonation model can be developed.

3. INTONATION MODELLING

Based on the physiological aspects described in section 2 and using prior work on intonation modelling, we derive a new way of decomposing intonation.

3.1. Prior work on intonation modelling

We are interested in intonation modelling for TTS. Describing all the models proposed to explicitly parameterise intonation would be a difficult task and we will only mention a few of the most popular ones.

The INTSINT model [6] aims to provide a transcription of intonation in a multi-lingual framework in an automatic way. An intonation curve is described as a sequence of target points, which are defined using the speaker's pitch range and/or the previous point as a reference. This model has been used for synthesis of French intonation [15].

The tilt model [7] is an evolution of the rise/fall/connection (RFC) model [16], which models intonation as a sequence of events. These events are described using three parameters: duration, amplitude and tilt, where the tilt parameter describes the shape of the event. These three parameters can be related to RFC parameters. They can be extracted automatically and the synthesis steps are straightforward. The main criticism on this model is the difficulty of predicting the parameters from linguistics.

Another interesting model is SFC [8] (Superposition of Functional Contours), which is a data driven model relying on metalinguistic information for synthesis. The model is superpositional as it is composed of several components trained using neural networks. The functional contours are extracted from a prosody rich corpus.

Finally, one of the most popular and that has been used a lot in the past is the command-response model [10], that will be described in more detail in Section 3.2.

All these models share a common distinctive feature which is the use of continuous F_0 contour. It was shown that using continuous F_0 improves the quality of synthesised intonation [17]. Therefore we use continuous F_0 contours and except if mentioned otherwise, F_0 will refer to a continuous curve in the rest of the paper.

3.2. Decomposition of F_0

Following previous work [18], Fujisaki defined the logarithm of F_0 as the sum of a baseline level, phrase components, and accent components [10].

This formulation, also known as the command-response model, assumes that the global shape of an utterance $\log F_0$ is generated from the response to phrase commands, while the local variations are accounted for by accent command response. This model assumes critically-damped second-order linear systems. Phrase components are then response of the system to impulsive driving forces, accent components are responses of the system to step driving functions.

A discrete-time version of Fujisaki's model was derived and a statistical model for the F_0 contours proposed [19]. Phrase and accent command pairs are modelled using an HMM with some particular constraints: a phrase component cannot occur while an accent command is still active, there cannot be overlap between accent commands. Using generative models such as HMM should allow for synthesis of plausible F_0 contours. This model was further improved to model duration of the step functions using substates [20].

The step response function to the accent command is equivalent to the impulse response to a train of phrase command like impulses (and we can relate it to the spikes sent from the brain to muscles). It is then possible to define all Fujisaki parameters in terms of impulse responses and therefore by responses of the type [10]:

$$G_{pi}(t) = \begin{cases} \alpha_i^2 t e^{-\alpha_i t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases} \quad (1)$$

Prom-on found that using higher order systems improves modelling [21]. Third order was found to perform better than second order, and fourth order to improve marginally compared with third order. Higher order critically damped systems lead to functions of the general gamma form, which has a convenient definition. Equation (1) can then be written as:

$$G_{k,\theta}(t) = \frac{1}{\theta^k \Gamma(k)} t^{k-1} e^{-t/\theta} \quad \text{for} \quad t \geq 0 \quad (2)$$

with $k = 2$ and $\theta = 1/\alpha$.

Based on these observations, we try to decompose F_0 using a set of kernel functions of the form (2). If higher order gamma shapes fit, we can infer that the underlying process is also higher order.

3.3. Using matching pursuit to decompose F_0

The matching pursuit algorithm [22] allows approximation of a signal $x(t)$ into a linear combination of so-called atoms, given a dictionary of kernel functions in the following way:

$$x(t) = \sum_{m=1}^{M} \sum_{i=1}^{I_m} \alpha_{m,i} \phi_m(t - \tau_{m,i}) + \epsilon(t) \quad (3)$$

where $\{\phi_1, ..., \phi_M\}$ is the dictionary, $\alpha_{m,i}$ is the gain of the instance i of kernel ϕ_m and $\tau_{m,i}$ its time delay, and ϵ is the residual error. Matching pursuit is a greedy algorithm which achieves this decomposition up to some error ϵ by operating iterative local optimisations. At each iteration, the algorithm computes the correlation between the signal and every atom from the dictionary, finds the most correlated atom and subtracts the weighted atom from the signal. These steps are applied on the residual signal until some accuracy threshold is reached.

As explained in section 2, we want to decompose the intonation contour using a two layer representation: a layer corresponding to a

phrase component, and a layer of local components corresponding to muscle responses. For that purpose, we decompose F_0 in a two pass process: first an iteration of matching pursuit is applied on the continuous F_0 contour to get the main component of the intonation. This is done with a dictionary of long atoms that can span more than the length of the utterance. The second step consists of applying matching pursuit on the residual obtained in the first step. The extracted atoms are then selected using a weighted root mean square error (wRMSE) criterion defined in 3.4.

The dictionary consists of atoms of the form of equation (2).

Our contour is then modelled as a base offset component, a phrase component, and the sum of gamma distribution shaped atoms:

$$\log F_0(t) = F_b + \alpha_p F_{0p}(t) + \sum_{i=1}^{Na} \alpha_i G_{k_i,\theta_i}(t - \tau_i) + \epsilon(t) \quad (4)$$

where $G_{k,\theta}$ and F_{0p} are defined in (2) (with $k = 2$ for F_{0p}) , α_p is the gain of the phrase component, and α_i and τ_i are the weights and the time offsets associated to the atoms $G_{k_i,\theta_i}(t)$, $\epsilon(t)$ is the residual. The phrase component is the same as in Fujisaki's model, except that θ can take different values.

The way matching pursuit decomposes F_0 does not allow to extract a train of impulses modelling an accent component from Fujisaki's model, as it will replace it by a bigger atom. However, Fujisaki developed the command-response model to model Japanese intonation, in which case it makes perfect sense due to the structure of the intonation produced in Japanese. Our model aims to be suitable for any language, and its physiological basis is a way to achieve it.

A summary of the procedure is given in Algorithm 1.

Algorithm 1 Atom decomposition

```
procedure ATOM DECOMPOSITION
    Extract F0, energy and POV from waveform.
    Subtract Fb = F0min.
    Extract F0p using matching pursuit and subtract it.
    Extract atoms using matching pursuit.
Loop:
    if wRMSE ≤ Threshold then
        goto End.
    else
        if Atom decreases wRMSE by more than 0.001 then
            Keep the atom and goto Loop.
        else
            Discard the atom and goto Loop.
        end if
    end if
End.
end procedure
```

3.4. Intonation similarity measures

Two methods were used to evaluate the perceptual distance of the intonation contour obtained with the atom decomposition process: the weighted root mean square error (wRMSE) and the weighted correlation coefficient, both introduced by Hermes [23]. The first is also used within the atom decomposition process.

The weighted RMS distance between the modeled and the originally extracted pitch contour, was calculated using (5). Here $w(i)$ and $p(i)$ are the weighing functions, i.e. the speech signal power and the probability of voicing (POV), $\hat{f}_0$ is the estimated version of the intonation contour f_0. Hermes introduced the subharmonic sumspectrum (SHS spectrum) [24] and used its maximum amplitude later for weighing RMS [24]. Rillard [25] and d'Allessandro [26] have suggested using the power of the speech signal instead, easing wRMSE calculation. We have opted for the latter, augmenting it with the POV calculated as detailed by Ghahremani [27]. Incorporating the POV in the weighing eliminates the need to hard threshold the POV to obtain voicing, making the whole approach more robust. It also brings the calculation of the wRMSE arguably closer to that originally proposed by Hermes [23].

$$R = \sqrt{\frac{\sum_i w(i)p(i)(\hat{f}_0(i) - f_0(i))^2}{\sum_i w(i)p(i)}} \quad (5)$$

It should be noted that the two pitch contours in (5) were calculated in semitones (ST), as in d'Allessandro's work [26], because of the possibility to relate the distance measures to perceptual research on intonation.

In line with (5), the weighted correlation was calculated using (6). Here $\hat{f}'_0(i)$ and f'_0 are the zero mean versions of the two pitch contours, and again the power of the speech signal was used for the weighting function augmented by the POV.

$$r = \frac{\sum_i w(i)p(i)\hat{f}'_0(i)f'_0(i)}{\sqrt{\sum_i w(i)p(i)\hat{f}'_0(i)^2 \sum_i w(i)p(i)f'_0(i)^2}} \quad (6)$$

4. EXPERIMENTAL FRAMEWORK

At the outset, aside from demonstrating the general utility of the atom based model, we have the opportunity to evaluate whether the model order suggested by Prom-on [21] is more suitable than that of Fujisaki [10]. Fujisaki's model is $k = 2$; if $k = 3$ or 4 turns out to be more efficient, we can conclude that Prom-on's higher order mechanism is more likely.

4.1. Data

The decomposition was run on a total of 60 sentences using speech from 6 different speakers and 3 different languages: English, French and German. For each language, a male (M) and a female (F) speaker were chosen: *rjs* (M), released for Blizzard Challenge 2010[1] and *slt* (F) for English [28], *Bernard* (M)[2] and *Isabelle Brasme* (F)[3] for French and *spid* (M) and *alzn* (F) for German [29].

4.2. Tools and settings

The Kaldi pitch tracker was used for F_0 and probability of voicing (POV) extraction [27]. We used 50ms frame length with 5ms frameshift for extraction. The matching pursuit toolkit (MPTK) was used to decompose intonation with our dictionaries [30].

4.3. Dictionaries

Following the MPTK dictionary requirements, we built some dictionaries that could be used to decompose F_0. All the atoms are based on eq. (2). Two sets of atoms were used: one set for the phrase component extraction, with $k = 2$ and $\theta = \{0.1, .., 0.8\}$ and one for local components. For the local components, several dictionaries were compared for $k = \{2, 3, ..., 7\}$, with $\theta = \{0.012, 0.014, ..., 0.8\}$.

[1] http://www.synsig.org/index.php/Blizzard_Challenge_2010

[2] https://librivox.org/a-lombre-des-jeunes-filles-en-fleur-by-marcel-proust-0905/

[3] https://librivox.org/la-princesse-de-cleves-by-madame-de-la-fayette/

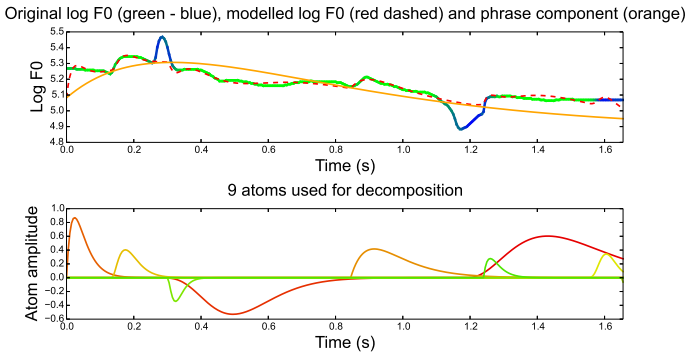

Fig. 1. *F_0 decomposition for the sentence* "I don't know why you're here at all" *by English female speaker* slt *using k=4. Top picture: For original F_0, POV is represented as highly voiced in green and unvoiced in blue, dashed red curve is reconstructed using atoms from bottom picture, orange curve is phrase component. Bottom picture: atoms extracted.*

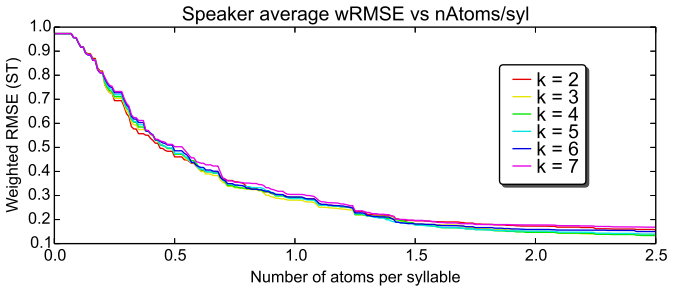

Fig. 2. *Weighted RMSE (in semitones) vs number of atoms per syllable for speaker* slt.

5. RESULTS

An example of decomposition is given in figure 1. The parts where original F_0 is green are highly probably voiced and the parts where the curve is blue are highly probably unvoiced. We can notice the influence of the weighted correlation-based atom selection, with strongly blue parts of the curve between 0.25 an 0.35 seconds and between 1.1 and 1.3 seconds. In these cases, the deviation of the curve is not considered by the model so atoms which would fit the curve are discarded, the dashed red curve is then smoothing F_0 in these unvoiced regions.

Figure 2 shows the average weighted root mean square error versus the number of atoms per syllable for the different k tested (from 2 to 7) for one example speaker (*slt*). We can notice that second and seventh orders perform worse when the WRMSE is getting lower. $k = 3, 4, 5, 6$ and were found to perform better than $k = 2, 7$ for all the speakers. An experiment was also done using all the k values in the same dictionary. As expected, using all the k's results in a slightly lower WRMSE with the same number of atoms per syllable, as the dictionary is bigger and offering more possibilities for decomposition. However, the greediness of the matching pursuit algorithm will optimise locally the decomposition, then sometimes choosing different k's to fit the curve. This was done to verify the underlying model, and we found that the majority of atoms selected are using $k = 4, 5, 6$. More generally speaking, increasing the order up to $k = 6$ is increasing the performance. In most of the cases, $k = 4$ was giving the best performance, in line with Prom-on's findings [21]. The high variance across sentences makes it difficult to strongly chose one k. Moreover there is no plausible reason to use several orders in the model, so we assume that using order 4 is more reasonable than 5 or 6 as they do not show significant differences.

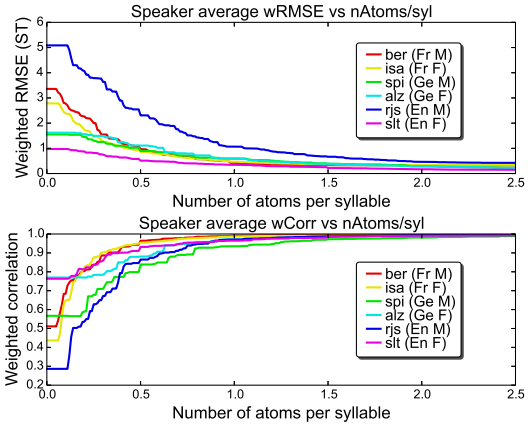

Fig. 3. *Weighted RMSE (in semitones) and correlation vs number of atoms per syllable for k = 4.*

Figure 3 shows the average WRMSE error versus the number of atoms per syllable and the average weighted correlation versus number of atoms per syllable for each speaker, using $k = 4$ which was found to perform best over all sentences, as we would expect, the global trend is an increasing correlation and a decreasing RMSE as the number of atoms per syllable increases. The desired accuracy can then be reached by selecting more atoms. Furthermore, using only one atom per syllable, the wRMSE and wCorr are close to 0 and 1, respectively. In languages like French and English, it makes sense to have one atom for each syllable, as they both have syllable-based stress. The language independent characteristic – at least for the languages under scrutiny – of our model is also demonstrated by similar results on the different languages.

Additionally, some informal listening tests revealed that resynthesising speech with the modelled F_0 was perceptually not different than the original F_0, even with small number of atoms. The listeners agreed that formal listening tests would not be more informative than the objective results.

6. CONCLUSION

Using previous work on intonation and on physiological aspects of intonation production, we proposed a model that decomposes intonation into gamma distribution-shaped atoms using a matching pursuit algorithm. The decomposition allows to model as precisely as desired a continuous intonation curve, and takes into account the voicing of speech. We found that higher order models were performing better than second order model. Only a few atoms are needed to achieve a level of accuracy that is not perceptually distinguishable from original intonation. The parameters associated to the atoms are easy to extract, and the approximation of intonation they provide is promising for prosody extraction and synthesis.

The atom dictionaries could be adapted to the data, using dictionary learning techniques. The model was originally envisaged as a means of connecting physiology to semantic meaning and some informal results have indeed showed a strong correlation between atoms and prosodic events. Linking atoms automatically to prosodic events is a matter of future work.

7. REFERENCES

[1] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *ICASSP*. IEEE, 2013, pp. 7962–7966.

[2] Heiga Zen and Andrew Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *ICASSP*. IEEE, 2014, pp. 3872–3876.

[3] Yao Qian, Yuchen Fan, Wenping Hu, and Frank K. Soong, "On the training aspects of deep neural network (DNN) for parametric (TTS) synthesis," in *ICASSP*. IEEE, 2014, pp. 3829–3833.

[4] Keiichi Tokuda, Heiga Zen, and Alan W Black, "An HMM-based speech synthesis system applied to english," in *Proc. of 2002 IEEE SSW*. IEEE, 2002, pp. 227–230.

[5] Janet Pierrehumbert, "Synthesizing intonation," *Journal of the Acoustical Society of America*, vol. 70, pp. 985–995, 1981.

[6] Daniel Hirst, Albert Di Cristo, and Robert Espesser, "Levels of representation and levels of analysis for the description of intonation systems," in *Prosody: Theory and experiment*, pp. 51–87. Springer, 2000.

[7] Paul Taylor, "Analysis and synthesis of intonation using the tilt model," *Journal of the Acoustical Society of America*, vol. 107, pp. 1697–1714, March 2000.

[8] Gérard Bailly and Bleicke Holm, "SFC: a trainable prosodic model," *Speech Communication*, vol. 46, no. 3, pp. 348–364, 2005.

[9] Antti Suni, Daniel Aalto, Tuomo Raitio, Paavo Alku, and Martti Vainio, "Wavelets for intonation modeling in HMM speech synthesis," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, August 2013, pp. 305–310.

[10] Hiroya Fujisaki and Shigeo Nagashima, "A model for the synthesis of pitch contours of connected speech," Tech. Rep., Engineering Research Institute, University of Tokyo, 1969.

[11] Greg Kochanski, Chilin Shih, and Hongyan Jing, "Quantitative measurement of prosodic strength in Mandarin," *Speech Communication*, vol. 41, no. 4, pp. 625–645, 2003.

[12] Helmer Strik, *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, October 1994.

[13] Ingo R. Titze and Daniel W. Martin, "Principles of voice production," *Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 1148, 1998.

[14] Hiroya Fujisaki, "Physiological and physical mechanisms for tone, accent and intonation," in *XXIII World Congress of the International Association of Logopedics and Phoniatrics*, Cairo, Egypt, 1995, pp. 156–159.

[15] Jean Véronis, Philippe Di Cristo, Fabienne Courtois, and Cédric Chaumette, "A stochastic model of intonation for text-to-speech synthesis," *Speech Communication*, vol. 26, no. 4, pp. 233–244, 1998.

[16] Paul Taylor, "The rise/fall/connection model of intonation," *Speech Communication*, vol. 15, no. 1, pp. 169–186, 1995.

[17] Kai Yu and Steve Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1071–1079, July 2011.

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

[19] Hirokazu Kameoka, Jonathan Le Roux, and Yasunori Ohishi, "A statistical model of speech F0 contours," in *Proceedings ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA)*, September 2010, pp. 43–48.

[20] Kota Yoshizato, Hirokazu Kameoka, Daisuke Saito, and Shigeki Sagayama, "Statistical approach to Fujisaki-model parameter estimation from speech signals and its quantitative evaluation," in *Speech Prosody*, 2012, pp. 175–178.

[21] Santitham Prom-on, Yi Xu, and Bundit Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *Journal of the Acoustical Society of America*, vol. 125, pp. 405–424, January 2009.

[22] Stéphane G. Mallat and Zhifeng Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[23] Dik J. Hermes, "Measuring the perceptual similarity of pitch contours," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 1, pp. 73–82, February 1998.

[24] Dik J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 257–264, 1988.

[25] Albert Rilliard, Alexandre Allauzen, and Philippe Boula de Mareüil, "Using dynamic time warping to compute prosodic similarity measures," in *Proceedings of Interspeech*, Florence, Italy, August 2011, pp. 2021–2024.

[26] Christophe d'Alessandro, Albert Rilliard, and Sylvain Le Beux, "Chironomic stylization of intonation," *Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1594–1604, March 2011.

[27] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *ICASSP*. IEEE, 2014, pp. 2513–2517.

[28] John Kominek and Alan W Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[29] Wolfgang J Hess, Klaus J Kohler, and Hans-Günther Tillmann, "The phondat-verbmobil speech corpus.," in *EUROSPEECH*, 1995.

[30] Sacha Krstulovìc and Rémi Gribonval, "MPTK: Matching pursuit made tractable," in *ICASSP*. IEEE, 2006, vol. 3, pp. 496–499.