Unified Prosody Model based on Atom Decomposition for Emphasis Detection

Branislav Gerazov[1], Aleksandar Gjoreski[1], Aleksandar Melov[1], Pierre-Edouard Honnet[2],
Zoran Ivanovski[1] and Philip N. Garner[2]
[1] Faculty of Electrical Engineering and Information Technologies,
University of Ss. Cyril and Methodius in Skopje, Macedonia
[2] Idiap Research Institute, Martigny, Switzerland
gerazov@feit.ukim.edu.mk, aleksandar@gjoreski.mk, aleksandar_melov@yahoo.com,
pehonnet@idiap.ch, mars@feit.ukim.edu.mk, pgarner@idiap.ch

***Abstract*—The prosody of the speech signal carries both linguistic and paralinguistic information. As such, there is a necessity of its modelling for the purpose of integrating it in speech technology systems. So far, there has been a multitude of proposed models focusing mainly on intonation, but a few also on energy and duration. The paper proposes an integrated approach to modelling the three dimensions of prosody through the use of atom decomposition techniques that we refer to as a Unified Prosody Model (UPM). The advantages of using such an integrated approach are illustrated in the task of emphasis detection, for which simple features are constructed based on the output of our UPM. A logistic regression classifier is trained and tested using these features and reaches an accuracy of 91%. This proof-of-concept algorithm illustrates the potential behind using the proposed UPM in improving prosody related speech research.**

Index Terms: prosody, intonation, energy, duration, modelling, emphasis

I. Introduction

Prosody is a multidimensional phenomenon comprising the intonation, energy, and duration contours of the speech signal. It carries both linguistic information, including: sentence structure, mode of enunciation, focus and contrast [1], as well as paralinguistic information, such as gender, age, emotions, and physiological state [2]. Because of this prosody is crucial in speech technology systems, especially in Text to Speech synthesis where it is necessary for generating natural speech output, but also in Automatic Speech Recognition of tonal languages. The importance of integrating prosody has been emphasized with the shift of scientific focus on the areas of Speech Emotion Recognition [3], emotional speech synthesis [4], and emphatic human-machine dialogue systems.

To facilitate its integration, various models have been developed for the different dimensions of prosody [5], most of them for modelling intonation. The intonation models have largely followed two general approaches: 1) modelling the pitch contour directly, and 2) modelling the underlying mechanisms, i.e. the physiology of pitch production. The former approach comprises a plethora of models, including ToBI [6], INSINT [7], Tilt [8], General Superpositional Model of Intonation [9], and SFC [10]. The physiological models differ from the surface models in that they directly or indirectly incorporate physiological constrains in the modelling process. These models include StemML [11] and qTA [12], as well as the most famous Fujisaki Command-Response (CR) model [13], which goes far into trying to model the pitch contour through the underlying laryngeal muscle activations [14].

In sharp contrast to the numerous intonation models, there have been only a few models for modelling duration and energy. These include the Klatt [15] and Sums-of-Products [16] duration models, and the functional data analysis (FDA) [17] and Legendre polynomials [18] based approaches to modelling the energy contour. Finally, an interesting approach was taken by Ward [19], who used Principal Component Analysis (PCA) to jointly model all the dimensions of prosody to extract underlying prosodic patterns.

Since almost all of the aforementioned models are designed for modelling a single dimension of prosody, practical systems must rely on different models for the different aspects of prosody. This can potentially raise inter-model compatibility issues, and could ultimately lead to a loss of interdimensional information. Moreover, some of the models are inherently ambiguous, e.g. ToBI [6], introducing errors in the analysis. Finally, most of the models, except the physiological intonation models, are distanced from the human process, reducing their usability in inferring higher levels of meaning from prosody. These problems have given rise to the trend of abandoning prosody models all together, and relying on machine learning algorithms trained for a specific task on “raw” prosody features [20]. However practical, this approach precludes a deeper understanding of the inner workings of prosody [19].

In this paper, we are proposing an integrated prosody modelling framework that we call a Unified Prosody Model (UPM). The model is based on an atom decomposition algorithm and draws on the physiology of prosody production, making it inherently language independent. The UPM is based on our work on a generalized CR intonation model and the Weighted Correlation based Atom Decomposition (WCAD) algorithm [21], [22], which offer improved consistency and physiological plausibility, as well as comparable performance to the standard CR model. We have previously extended the atom decomposition approach towards modelling the energy contour [23], and in this paper, we further this extension towards creating a unified modelling framework for prosody. This unified modelling framework is advantageous in fields relying on the multidimensional analysis of prosody, because it allows for a consistent framework of describing the different dimensions of prosody.

As an example case-study we have chosen the task of emphasis detection, which is gaining importance because of its use in speech-to-speech translation systems and human-machine dialogue systems. We have chosen this task on the merit that emphasis has been shown to be indicated by all of the dimensions of prosody [24]: intonation [25], energy [26], [27], [28], and duration [29]. The results show that our UPM is a good foundation for building a state-of-the-art emphasis detection algorithm.

II. Integrated prosody model

The UPM is based on the decomposition of the three prosody contours into a set of elementary gamma distribution based atoms. The decomposition process of the UPM is based on a matching pursuit framework [30] and comprises two general steps: 1) extract a global phrase atom, and 2) iteratively extract smaller local atoms. The phrase atom has a different physiological interpretation for each of the dimensions of prosody.

The elementary atoms (1) are designed to capture the physiological muscle response to an elementary impulse excitation, i.e. a muscle twitch. The gamma distribution used to generate them is a higher order generalization of the 2nd order spring-mass-damper system used in the standard CR model [14], [31]. This is physiologically plausible when using more complex muscle models, based on the 3rd order Hill model [32], such as our recently proposed agonist-antagonist pitch production (A2P2) model [33].

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

The UPM atom decomposition is outlined in Algorithm 1.[1] During initialization the algorithm extracts: a continuous pitch estimate f_0 [34], the energy e, the syllable duration as a z_{score}, the probability of voicing (POV) p [34] used to calculate a weighting function w, and the start t_s and end t_e of phonation. For each of the three prosody contours c a phrase atom a_{phrase} is extracted by selecting the atom that maximizes a chosen cost function within a range determined using t_s and t_e. The amplitude of the atom is then calculated using standard correlation, and is subtracted from c to obtain c_{diff}. Next, local atoms a_{local} are iteratively extracted from c_{diff} using the cost function. Each new atom's amplitude is again calculated using the standard correlation and is subtracted from c_{diff}. Local atom extraction ends when either the amplitude of the atoms goes below a set threshold.

An example UPM decomposition is shown in Fig. 1. The top plot shows the audio signal waveform with the annotations in IPA, the 2nd shows the pitch contour and the corresponding pitch phrase atom, with the 3rd showing the extracted pitch local atoms. The 4th and 5th plot show the energy contour and the phrase and local atoms. Finally, the bottom two plots show the duration contour and the extracted atoms.

Algorithm 1 Integrated Prosody Model atom decomposition.

```
procedure UPM Atom Decomposition
    Extract $prosody = [f_0, e, z_{\text{score}}]$
    Extract $p$, calculate $w = e \cdot p$
    Extract $t_s$ and $t_e$ of phonation
    for $c$ in $prosody$ do
        Extract $a_{\text{phrase}}$ using $cost()$
        Calculate $a_{\text{phrase}}$ amplitude using $corr()$
        $c_{\text{diff}} = c - a_{\text{phrase}}$
        repeat
            Extract $a_{\text{local}}$ using $cost()$
            Calculate $a_{\text{local}}$ amplitude using $corr()$
            $c_{\text{diff}} = c_{\text{diff}} - a_{\text{local}}$
        until $max(a_{\text{local}}) < a_{\text{thresh}}$
    end
```

A. Modelling intonation

The decomposition of the pitch contour is entirely based on our WCAD[2] algorithm [22]. The decomposition of the pitch contour takes place in the $\log f_0$ domain, and it differs from the decomposition of the energy and duration contours in that it uses the weighted correlation (WCORR) (2) calculated using the weight w, as a cost function due to its perceptual significance [35]. Here f_0 is the reference pitch contour, $\hat{f}_0$ is the modelled one.

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

The phrase atom in the pitch contour corresponds to the global component of the subglottal pressure P_{sb} which decreases in the process of phonation, gradually lowering the pitch [36]. Due to the lack of the rise portion of this global component, its fall is only modelled. This is done by placing the maximum of the atom kernel (1) at the start of phonation t_s, and finding the θ_f that maximizes the WCORR within the range t_s to $t_e - t_{\text{off}}$, where t_{off} is an offset introduced to eliminate possible phrase final accents from the analysis.

B. Modelling energy

The decomposition of the energy contour follows the general UPM algorithm, with some differences [23]. Since the global component of the energy contour is also due to fall in P_{sb}, the extracted pitch phrase atom, is used to scale the energy contour. The extracted energy phrase atom is used to scale the energy contour using (3), which normalizes, inverts and adds a DC offset to it [23]. In this way the energy contour towards the end of phonation is amplified, while maintaining its original values near the start of phonation, as seen in the 3rd plot of Fig. 1. This approach has been chosen over reducing the starting energy levels in order to keep the energy contour above 0. The scaled energy contour e_{scaled} is then decomposed iteratively using the standard correlation as a cost function.

$$e_{\text{scaled}} = e \cdot \left(1 - \frac{a_{\text{phrase}}}{max(a_{\text{phrase}})} + 1\right) \tag{3}$$

[1] The UPM implementation code is available on gitHub at https://github.com/dipteam/upm

[2] The WCAD implementation code is available on gitHub at https://github.com/dipteam/wcad

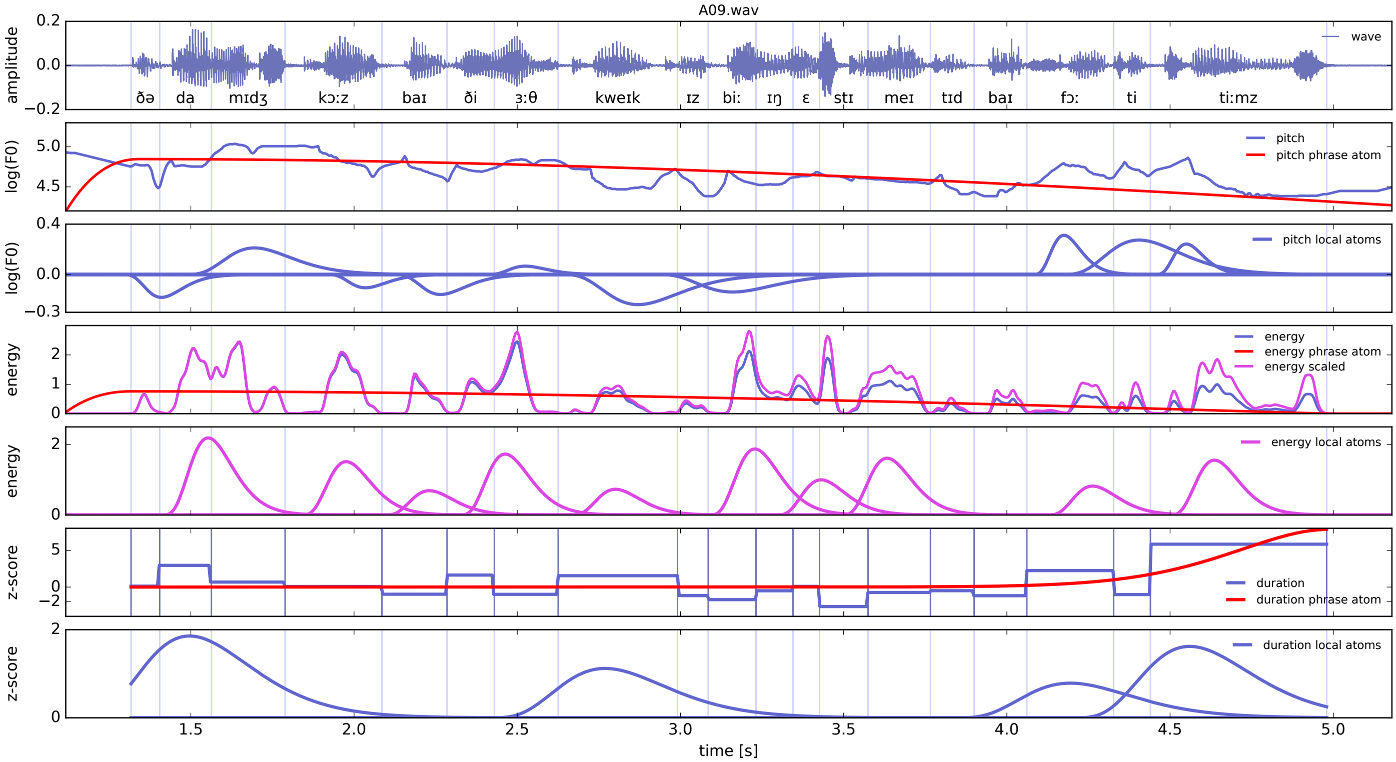

Fig. 1. *Example UPM decomposition of the utterance: "The damage caused by the earthquake is being estimated by forty teams."*

C. Modelling duration

The duration contour modelled with the UPM is represented by a step function whose amplitude for each syllable is equal to the z-score of its duration [29]. The z-score is calculated by summing the z-scores for each of the phones constituting the syllable according to (4). Here P is the number of phones in the syllable, z_i are their z-scores, t_i is the duration of the i^{th} phone, μ_i its average duration, and σ_i the standard deviation of its duration.

$$z_{syllable} = \sum_{i=1}^{P} z_i = \sum_{i=1}^{P} \frac{t_i - \mu_i}{\sigma_i} \tag{4}$$

The global component of the duration contour is marked by a phrase final rise corresponding to the phrase final lengthening of the last syllables [37]. This lengthening is due to the relaxation of the articulators at the end of the utterance which effects the slowing down of their dynamics, and thus increasing the duration of articulation. Since this phenomenon differs from the one governing the phrase component in the pitch and energy contours, its extraction is done independently by fitting the fall part of the gamma kernel, which was mirrored left-to-right. Similar to the phrase atom extraction from the pitch contour, the location of the atom's maximum is fixed to the end of phonation t_e and the θ_f is selected that maximizes the cost function, which for the duration contour is the correlation function. An example of this can be seen in the 6th plot in Fig. 1.

III. Emphasis detection

The communication of emphasis in the speech signal can rely on using any or all of the dimensions of prosody. In this sense the integrated modelling framework offered by the UPM is well suited for the task of automatic emphasis detection. An example of this is shown in Fig. 2, where the duration atoms point to which of the energy atoms correspond to emphasis. Our hypothesis is that the UPM will be able to capture the increased prosodic effort speakers make when they want to emphasise a word through the increased dynamics in the atom amplitudes.

A. Experiments

To assess the usability of UPM for the task of emphasis detection we will use a subset of the English part of the multilingual SP2 Speech Corpus[3] [38] created under the SP2 project [39]. The subset comprises 4 sets of 10 utterances from the English speaker which have deliberately placed emphasis:

A. Emphasis on one word in the utterances,
B. Emphasis at the start of the utterances,
C. Contrastive emphasis, and
D. Contrastive emphasis in a question.

This amounted to 381 words in total of which 66, or 17% were emphasised. For the analysis we first use the UPM to decompose the prosody of these 40 utterances into elementary atoms. Then we construct a feature vector using the maximum atom amplitudes for each dimension of prosody per word. We feed the feature vectors to a logistic regression machine learning algorithm. For estimating the classifier performance we use a 5-fold cross-validation assessment, in which the data is randomly divided into 5 subsets, of which repeatedly one is

[3] The SP2 Speech Corpus can be downloaded from https://github.com/SP2-Consortium/SP2-Speech-Corpus

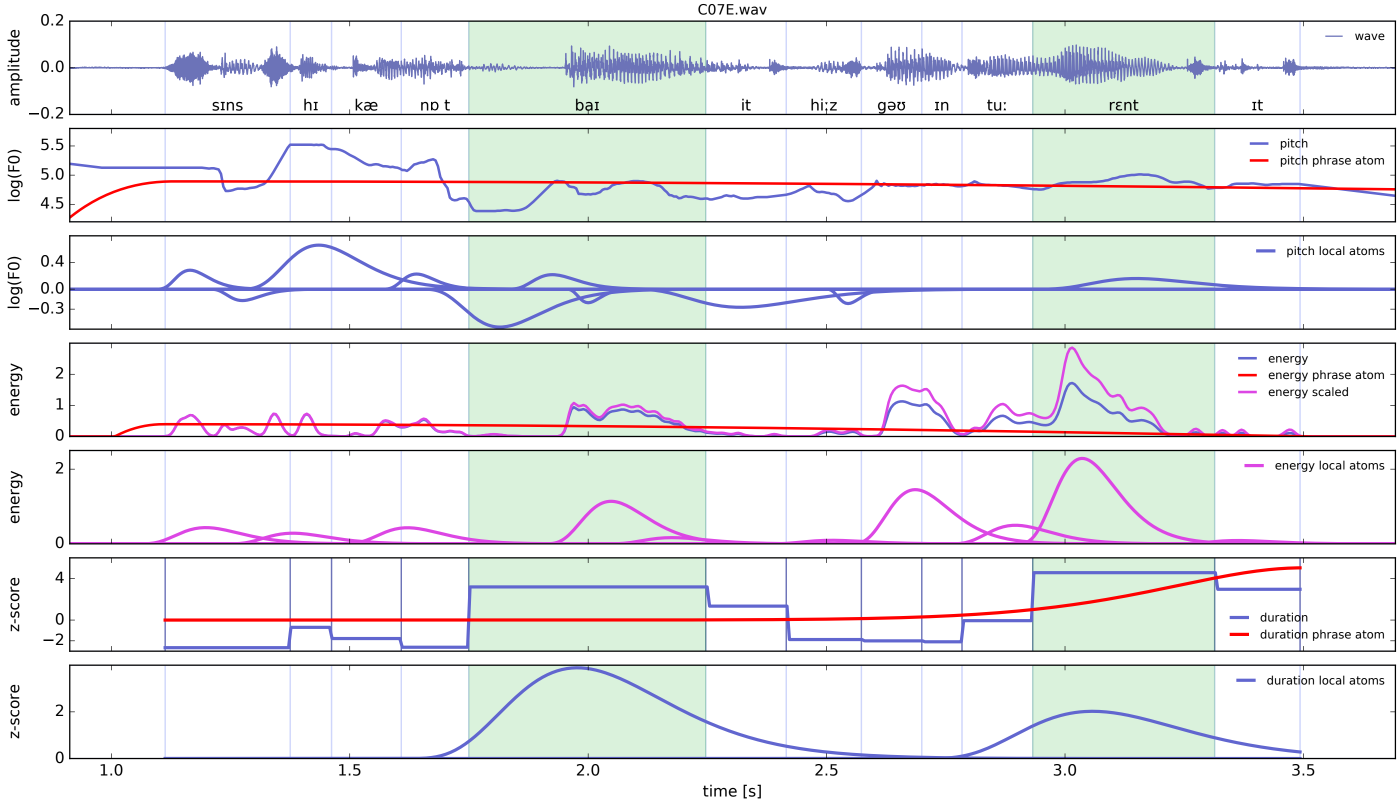

Fig. 2. *Example UPM decomposition for the purpose of emphasis detection for an utterance with contrastive emphasis: "Since he cannot* ***buy*** *it, hes going to* ***rent*** *it."*

used for testing and the others for training, giving a 20/80% test/train ratio.

B. Results

The distribution of the atom amplitude feature vectors is shown in Fig. 3, where emphasised words are marked yellow, and non-emphasised word violet. We can see that although there is a large amount of overlap, the atom amplitudes of the emphasised words have a significantly larger variance, as was hypothesised. The average accuracy was calculated using A 5, where $\hat{y}$ are the predicted values, and y are the true values. The accuracy of the classifier assessed for a 5-fold cross-validation on the whole dataset was 0.92 ±0.03%. On the other hand, the same analysis the corresponding values for the precision 6 and recall 7 are 0.87 ±0.14% and 0.66 ±0.19%, respectively, where TP are the correctly detected "true" positives, FP are the falsely detected positives, and P is the total number of positives in the data. This performance of the classifier can be seen in the precision-recall curve for varying decision thresholds in Fig. 4.

$$A(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N-1} 1(\hat{y}_i = y_i) \tag{5}$$

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{P} \tag{7}$$

For comparison the state-of-the-art results of our previous algorithm that relied solely on the atom decomposition of the energy contour [23], and which was based on an adaptation of an approach based on probabilistic amplitude demodulation (PAD) [28], gave a precision and recall of 0.8, 0.8 and 1, 1, however only on a subsets A and B from the data, and by detecting a single emphasis per utterance.

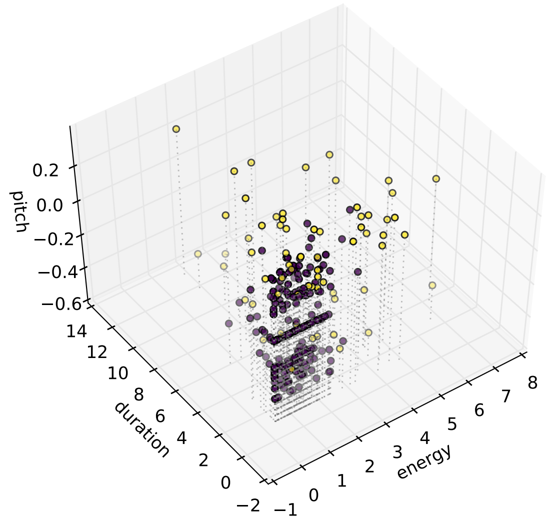

Fig. 3. *Distribution of the maximum atom amplitudes for each of the 381 emphasised (yellow) and non-emphasised (violet) words used in the experiments for each dimension of prosody.*

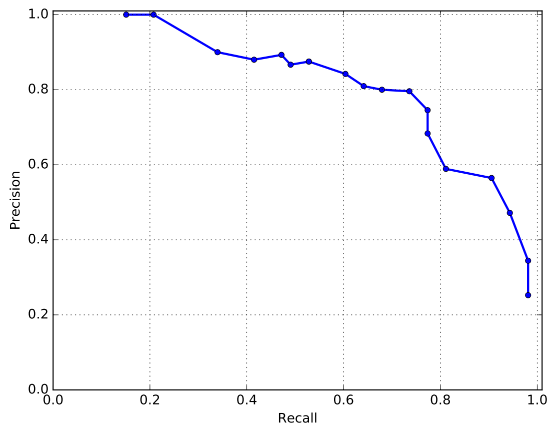

Fig. 4. *Precision-recall curve for the logistic regression classifier.*

IV. Conclusions

The proposed Unified Prosody Model is an integrated framework for the analysis of the phenomenon of prosody in all three of its dimensions. It uses a matching pursuit approach to first extract a global phrase atom from each of the three contours, and then to decompose them into local elementary atoms. The approach draws on the physiology of prosody production, and is thus inherently language independent. The integrated approach to prosody analysis is advantageous in that it facilitates the analysis of the three dimensions of prosody in a consistent way. We have shown the usefulness of the model for the task of emphasis detection, in which our UPM captured the hypothesised increased variability in the emphasised words.

A logistic regression classifier was trained using simple features extracted from the output of our UPM. The classifier showed high accuracy in its performance, but failed to show good precision and recall. It can, however, predict any number of emphasised words in the utterance, not just one. In this regard it outperforms previous state-of-the-art approaches. These proof-of-concept results are promising, and more advanced algorithms for emphasis detection should be explored based on the proposed UPM. Moreover, we believe that the UPM will be useful for advancing any field of speech research that relies on the analysis of prosody.

V. Acknowledgements

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

References

[1] Anne Cutler, Delphine Dahan, and Wilma Van Donselaar, “Prosody in the comprehension of spoken language: A literature review,” *Language and speech*, vol. 40, no. 2, pp. 141–201, 1997.

[2] Björn Schuller and Anton Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*, John Wiley & Sons, 2013.

[3] Thurid Vogt, Elisabeth André, and Johannes Wagner, “Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation,” in *Affect and emotion in human-computer interaction*, pp. 75–91. Springer, 2008.

[4] Felix Burkhardt and Nick Campbell, “Emotional speech synthesis,” *The Oxford Handbook of Affective Computing*, p. 286, 2014.

[5] Jan van Santen, Taniya Mishra, and Esther Klabbers, “Prosodic processing,” in *Springer Handbook of Speech Processing*, pp. 471–488. Springer, 2008.

[6] Kim EA Silverman, Mary E Beckman, John F Pitrelli, Mari Ostendorf, Colin W Wightman, Patti Price, Janet B Pierrehumbert, and Julia Hirschberg, “Tobi: a standard for labeling english prosody.,” in *ICSLP*, 1992, vol. 2, pp. 867–870.

[7] 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.

[8] Paul Taylor, “Analysis and synthesis of intonation using the tilt model,” *The Journal of the acoustical society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[9] Jan PH Van Santen and Bernd Möbius, “A quantitative model of f0 generation and alignment,” *IntonationAnalysis, Modelling and Technology*, pp. 269–288, 2000.

[10] Gérard Bailly and Bleicke Holm, “Sfc: a trainable prosodic model,” *Speech Communication*, vol. 46, no. 3, pp. 348–364, 2005.

[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] Santitham Prom-On, Yi Xu, and Bundit Thipakorn, “Modeling tone and intonation in mandarin and english as a process of target approximation,” *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 405–424, 2009.

[13] Hiroya Fujisaki, “A model for synthesis of pitch contours of connected speech,” *Annual Report, Engineering Research Institute, University of Tokyo*, vol. 28, pp. 53–60, 1969.

[14] Hiroya Fujisaki, “The roles of physiology, physics and mathematics in modeling prosodic features of speech,” in *Proc. of Speech Prosody*, 2006.

[15] Dennis H Klatt, “Synthesis by rule of segmental durations in english sentences,” *Frontiers of speech communication research*, vol. 1, pp. 287–300, 1979.

[16] Jan PH Van Santen, “Exploring n-way tables with sums-of-products models,” *Journal of mathematical psychology*, vol. 37, no. 3, pp. 327–371, 1993.

[17] Juan Pablo Arias, Carlos Busso, and Nestor Becerra Yoma, “Energy and f0 contour modeling with functional data analysis for emotional speech detection.,” in *INTERSPEECH*, 2013, pp. 2871–2875.

[18] Najim Dehak, Pierre Dumouchel, and Patrick Kenny, “Modeling prosodic features with joint factor analysis for speaker verification,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2095–2103, 2007.

[19] Nigel G Ward, “Automatic discovery of simply-composable prosodic elements,” in *Speech Prosody*, 2014, vol. 2014, pp. 915–919.

[20] Elizabeth Shriberg and Andreas Stolcke, “Direct modeling of prosody: An overview of applications in automatic speech processing,” in *Speech Prosody 2004, International Conference*, 2004.

[21] Pierre-Edouard Honnet, Branislav Gerazov, and Philip N. Garner, “Atom decomposition-based intonation modelling,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, Australia, April 2015, IEEE.

[22] Branislav Gerazov, Pierre-Edouard Honnet, Aleksandar Gjoreski, and Philip N. Garner, “Weighted correlation based atom decomposition intonation modelling,” in *Proceedings of Interspeech*, Dresden, Germany, September 2015.

[23] Aleksandar Gjoreski, Branislav Gerazov, and Zoran Ivanovski, “Atom-decomposition based analysis for the purpose of emphatic word detection,” in *XII International Conference ETAI*, Ohrid, Macedonia, September 2015.

[24] Jacques Terken and Dik Hermes, “The perception of prosodic prominence,” *Prosody: Theory and experiment. Studies presented to Gösta Bruce. Dordrecht: Kluwer*, pp. 89–127, 2000.

[25] D Robert Ladd and Rachel Morton, “The perception of intonational

emphasis: continuous or categorical?," *Journal of Phonetics*, vol. 25, no. 3, pp. 313–342, 1997.

[26] Mattias Heldner, Eva Strangert, and Thierry Deschamps, "A focus detector using overall intensity and high frequency emphasis," in *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, 1999, vol. 2, pp. 1491–1493.

[27] Ana Stojkovic, Branislav Gerazov, and Zoran Ivanovski, "Emphatic word detection based on relative phoneme energies within syllables," in *XII International Conference ETAI*, Ohrid, Macedonia, September 2015.

[28] Milos Cernak and Pierre-Edouard Honnet, "An empirical model of emphatic word detection," in *Proceedings of Interspeech*, Dresden, Germany, September 2015.

[29] Aleksandar Melov, Branislav Gerazov, and Zoran Ivanovski, "Emphatic word detection based on syllable durations," in *XII International Conference ETAI*, Ohrid, Macedonia, September 2015.

[30] 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.

[31] Branislav Gerazov and Philip N. Garner, "An investigation of muscle models for physiologically based intonation modelling," in *Proceedings of the 23rd Telecommunications Forum*, Belgrade, Serbia, November 2015, pp. 468–471.

[32] Vladimir Zatsiorsky and Boris Prilutsky, *Biomechanics of skeletal muscles*, Human Kinetics, 2012.

[33] Branislav Gerazov and Philip N. Garner, "Agonist-antagonist pitch production model," in *SPECOM*, Budapest, Hungary, August 2016. (accepted for publication).

[34] 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.

[35] 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.

[36] 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.

[37] Aleksandar Melov, Branislav Gerazov, and Zoran Ivanovski, "Towards extracting the global component from the syllable duration contour for emphatic word detection," in *3rd International Acoustics and Audio Engineering Conference TAKTONS*, Novi Sad, Serbia, November 2015.

[38] Milan Sečujski, Branislav Gerazov, Tamás Gábor Csapó, Vlado Delić, Philip N. Garner, Aleksandar Gjoreski, David Guennec, Zoran Ivanovski, Aleksandar Melov, Géza Németh, Ana Stojković, and György Szaszák, "Design of a speech corpus for research on cross-lingual prosody transfer," in *SPECOM*, Budapest, Hungary, August 2016. (accepted for publication).

[39] György Szaszák, Tamás Gábor Csapó, Philip N. Garner, Branislav Gerazov, Zoran Ivanovski, Géza Németh, Bálint Tóth, Milan Sečujski, and Vlado Delić, "The SP2 SCOPES project on speech prosody," in *Proceedings of the DOGS - Digital speech and image processing conference*, Novi Sad, Serbia, October 2014, pp. 2–10.