Weighted Correlation based Atom Decomposition Intonation Modelling

Branislav Gerazov[1], *Pierre-Edouard Honnet*[2], *Aleksandar Gjoreski*[1], *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, {pehonnet, pgarner}@idiap.ch, aleksandar@gjoreski.mk

Abstract

Intonation modelling is an integral part of text-to-speech systems from their very beginnings. This has led to the proliferation of various intonation models, each with its own relative strengths and weaknesses. Only a few of these intonation models are based on physiology, despite the advantage that such models are language independent. We propose a new intonation model inspired by the physiology of intonation production, which is based on decomposing the F_0 contour into elementary atoms. The model, named the Weighted Correlation Atom Decomposition model (WCAD), is a generalisation of the command response (CR) model and has the advantage of having a simple parameter extraction method. The decomposition process follows a matching pursuit approach based on using the perceptually relevant weighted correlation as a cost function. The results have affirmed the plausibility of using the WCAD model to model F_0 contours across different languages and speakers. The results have also shown that the WCAD model has good comparative performance to the CR model, giving it practical importance.

Index Terms: intonation model, physiology, matching pursuit, weighted correlation

1. Introduction

Current state of the art text-to-speech (TTS) synthesis systems are able to produce speech with reasonably good quality. However, one issue in TTS is the still unsatisfactory prosody of the synthetic speech. As an important part of prosody, intonation is one of the research areas that remain open in TTS. Many different approaches to intonation modelling can be found in the literature, but only a few of them are trying to model the production of intonation [1, 2, 3].

The command-response (CR) model [1] is the most popular in this category and more generally one of the most popular intonation models. It defines intonation by looking at the physiological process behind its production. Relating the vocal folds' tension with the activity of the muscles ruling them, Fujisaki showed that the fundamental frequency can be decomposed in several additive components in the log domain [4]. In this model, two different types of components are credited to the translation and rotation of the cricothyroid (CT) muscle.

Strik [5] identified more muscles at work in the production of F_0, as well as the influence of the subglottal pressure P_{sb}. In our previous work, we investigated a generalisation of the CR model that would account for more than 2 types of movements influencing the F_0 production [6]. Following some work on modelling intonation using the CR model [7, 8], we replaced the step functions used for local components in the CR model by impulses. By defining all the components of the model as impulse responses to a critically damped system, we argue that these components could be linked to the response of the muscles involved in intonation generation. The parameters of this model also present the advantage of being easy to extract using the matching pursuit algorithm [9], followed by a selection of the relevant atoms using a weighted root mean square distance.

In this paper, we present further development of our model by directly extracting relevant atoms by using weighted correlation; the perceptual relevance of the weighted correlation is discussed. We also present a different definition for the global component of intonation. The paper is structured as follows: Section 2 presents the weighted correlation-based decomposition method, Section 3 describes the experiments, Sections 4 and 5 give results and conclude the paper.

2. Weighted Correlation based Atom Decomposition

The Weighted Correlation based Atom Decomposition (WCAD) algorithm is based on the integration of the modified version of the perceptually relevant weighted correlation (WCORR) measure [10] as a cost function in the matching pursuit framework [9]. The algorithm is an improved and more compact version of the algorithm we presented in our previous work [6], which was based on using the weighted RMS error (WRMSE) to filter out the perceptually irrelevant atoms output by the matching pursuit algorithm. The WCAD algorithm also introduces a novel physiologically inspired type of atom for representing the global, i.e. phrase component in the F_0 contour.

2.1. Weighted Correlation

The weighted RMS and the weighted correlation measures were both first introduced in the work of Hermes [10]. In his work, Hermes found that the WCORR measure had the best correlation (0.67) with the auditory dissimilarity ratings of five experienced phoneticians. This is a solid correlation, having in mind that the interexpert agreement was found to be 0.65, in the same work. Moreover, Hermes found approximate thresholds that can be used to classify the perceptual similarity of two intonation contours using the WCORR, given in Table 1.

The weighted correlation (WCORR) introduced by Hermes [10] is calculated according to (1). Here f_0 is the reference F_0, $\hat{f}_0$ is the modelled F_0, i.e. its reconstruction, f_{0m} and $\hat{f}_{0m}$ are their respective means, and $w(i)$ is the weighting function. Originally, 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 [11].

$$r = \frac{\sum_i w(i)(\hat{f}_0(i) - \hat{f}_{0m})(f_0(i) - f_{0m})}{\sqrt{\sum_i (w(i)(\hat{f}_0(i) - \hat{f}_{0m})) \sum_i (w(i)(f_0(i) - f_{0m}))}} \tag{1}$$

In our WCAD algorithm, we use a modified version of the WCORR, given in 2, in which we alter the originally proposed WCORR in three ways. 1) We do not subtract the mean of the F_0 contours, as our model does not have a static component, and we need the extracted atoms to build up the F_0 contour from scratch. 2) We use the logarithm of the F_0, instead of the equivalent rectangular bandwidth (ERB) scale [12], because it is both traditionally used in intonation modelling research [1], and is also essentially equivalent to the use of semitones in perceptual intonation studies [13, 14]. 3) We redefine the weight according to (3), abandoning the deprecated SHS spectrum, and in accordance with newer trends in perceptual intonation studies [14, 13]. Here, $p(i)$ is the probability of voicing (POV) of frame i as defined by Ghahremani et al. [15], and $e(i)$ its energy. The use of a continuous POV estimate, instead of using a binary thresholded one [13], eliminates the use of hard thresholds, making the weighting function more robust.

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

$$w(i) = p(i)e(i) \tag{3}$$

2.2. Phrase atoms

In our previous work [6], we introduced the use of general gamma form atoms, defined according to (4), as the building blocks of the F_0 contour. This is based on a higher-order extension of the critically-damped second-order linear systems [16] that account for the phrase and accent commands in the original command-response model [1].

$$G_{k,\theta}(t) = \frac{1}{\theta^k \Gamma(k)} t^{k-1} e^{-t/\theta} \quad \text{for} \quad t \geq 0 \tag{4}$$

The problem with the use of gamma distribution shaped function to model the phrase atom is that the high θ-s needed to produce atoms with a sufficiently gradual fall, also stretched out the rise of the atoms and their peaks. This is in contrast to the qualitative shape of the global component of the subglottal pressure P_{sb}, as seen in the measurements of Strik [5]. There, the global component has a steeper rise with a relatively sharp peak at the start of phonation, which is followed by a lengthy fall. This reflects the initial buildup of P_{sb} that precedes speech, and its timely release for the purpose of sustaining phonation. We seek to capture this observed quality of the global P_{sb} component using a modified definition of the phrase atom, based on the concatenation of two gamma distribution functions:

$$G_{k,\theta_r,\theta_f}(t) = \begin{cases} \frac{1}{\theta_r^k \Gamma(k)} t^{k-1} e^{-t/\theta_r} & \text{for} \quad 0 \leq t \leq t_m \\ \frac{1}{\theta_f^k \Gamma(k)} (t - t_o)^{k-1} e^{-(t-t_o)/\theta_f} & \text{for} \quad t > t_{rm} \end{cases} . \tag{5}$$

Here, θ_r and θ_f are the two time constants for the rise and fall of the phrase atom, t_{rm} is the time instant of the phrase atom peak, and t_o is an offset meant to compensate for the difference between t_{rm} and the maximum of the fall function t_{fm}:

$$t_o = t_{rm} - t_{fm}. \tag{6}$$

Table 1: Weighted correlation thresholds for five perceptual similarity categories of two F_0 contours found by Hermes [10].

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

Algorithm 1 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 θ_f for $phrase\ atom$ at position t_s giving max WCORR for $t_s \leq t \leq t_e - t_{off}$.
    Calculate $phrase\ atom$ amplitude using correlation.
    $f_{diff} = f_0 - phrase\ atom.$
    $f_{recon} = phrase\ atom.$
Loop:
    Extract $local\ atom$ giving max WCORR with f_{diff} for $t > t_s$.
    Calculate $local\ atom$ amplitude using correlation.
    Increment $atom\ count.$
    $f_{diff} = f_{diff} - local\ atom.$
    $f_{recon} = f_{recon} + local\ atom.$
    if $\text{WCORR}_{\text{norm}}$ of $f_{recon} > \text{WCORR}_{\text{norm}}\ thresh$ **then**
        goto *End*.
    else
        goto *Loop*.
End.

2.3. WCAD model extraction

The Weighted Correlation Atom Decomposition (WCAD) algorithm is outlined in Algorithm 1. First, the algorithm extracts the energy e, f_0 and POV p, and calculates the weighting function w. Next, the start and end times of phonation, t_s and t_e, are estimated by finding the time instants when the energy e crosses a start threshold value T_s, and when it finally goes below a terminal threshold value T_e. The phrase atom peak is then positioned at t_s, and we find the θ_f that maximizes the WCORR (2) within a range between t_s and $t_e - t_{off}$. Here, t_{off} is an offset time introduced to leave out a possible phrase-final fall and rise in the F_0 contour from the phrase atom fitting. Also, we use a fixed value for θ_r due to the consistency in rise times across the utterances observed in Strik's measurements [5]. The amplitude of the phrase atom is calculated using the standard correlation, and is subtracted from f_0 to obtain f_{diff}. The phrase atom is also used to initialise the F_0 reconstruction f_{recon}.

In the second part of the atom decomposition, local atoms are extracted from f_{diff} using the WCORR, by selecting the atom that maximises it at each iteration. The amplitude of the extracted atoms is again calculated using the standard correlation, and they are subtracted from f_{diff}, and added to f_{recon}. Local atom extraction ends when either a) the reconstruction f_{recon} reaches a selected $\text{WCORR}_{\text{norm}}$ threshold, or b) when the chosen maximum number of atoms is reached. Here,

$\text{WCORR}_{\text{norm}}$ includes the zero-mean normalisation of the two F_0 contours as in (1), which allows us to use the WCORR perceptual thresholds from Table 1 as a stopping criterion. We argue that this is plausible, because the weight used by the WCAD algorithm essentially captures the same information as the originally proposed SHS. A formal proof of this is beyond the scope of the paper.

3. Experiments

We have designed two experiments to assess the plausibility of the introduced Weighted Correlation Atom Decomposition algorithm, and to compare its performance with a state of the art implementation of the standard CR model, as it is a generalised CR model.

Experiment 1. Our first goal is to analyse how well the WCAD algorithm models the F_0 contour. To assess this, we will analyse how much the addition of each local atom increases the $\text{WCORR}_{\text{norm}}$ between the original and modelled F_0 contours. We expect the WCORR to increase rapidly as the initial large local atoms are added and saturate at the optimal number of atoms per syllable. We also evaluate the number of atoms per syllable necessary to match the perceptual WCORR thresholds from Table 1. To extract the F_0 and the POV we will use the pitch tracker implemented in Kaldi [15]. It generates a continuous F_0 contour through the use of interpolation and smoothing.

Experiment 2. In order to compare the performance of the WCAD algorithm with the CR model, we use Mixdorff's CR parameter extraction tool [17]. Because this tool only outputs the final optimised CR model parameters, the calculated $\text{WCORR}_{\text{norm}}$ for the modelled F_0 will be plotted as single points in the WCORR – atoms/syl plain, and then compared to the results obtained with the WCAD algorithm. The average WCORR and number of commands per syllable, obtained with the CR model per speaker, will also be used in the comparison.

3.1. Database

The experiments were run on the same dataset used in our previous work [6]. This dataset contains a total of 60 utterances, and comprises recordings of 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 [18], *Bernard* (M)[2] and *Isabelle Brasme* (F)[3] for French and *spid* (M) and *alzn* (F) for German [19].

3.2. 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. It's reasonable to suppose that these are speaker dependent, but for the purpose of this paper we pull them together and assume speaker independence.

To determine the start of phonation t_s, we chose a threshold value T_s for the normalised energy of 0.5. For the end of phonation t_e, the threshold T_e was lowered to 0.1, because of the gradual decrease of energy towards the end of an utterance. 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.

[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/

Table 2: Number of atoms/syllable needed on average to reach a chosen perceptual similarity category, for each of the speakers.

Cat.	en M	en F	fr M	fr F	ge M	ge F	Avg.
1	0.71	0.79	0.38	0.49	0.75	0.83	0.66
2	0.48	0.42	0.29	0.30	0.47	0.53	0.41
3	0.34	0.27	0.19	0.19	0.32	0.39	0.28
4	0.26	0.17	0.14	0.12	0.24	0.24	0.19

The θ_r for the rising part of the phrase atoms was fixed at 0.5. The range for the θ_f of the phrase atoms was set to 0.1 - 10, and for the θ of the local atoms to 0.01 - 0.05. These two ranges give the needed atom variability in the WCAD algorithm. The values k in the atom gamma distribution shaped function 4 was set to 6, as it was found to have a slightly better overall performance than the k of 4 used in our previous work [6].

4. Results

Example results of the Weighted Correlation based Atom Decomposition algorithm are given in Fig. 1 for an utterance taken from the male French speaker in our database. The top panel shows the original F_0 contour, the extracted phrase atom and the reconstructed F_0 by our model. The local atoms that compose this contour are given in the middle panel. Finally, the bottom panel shows the energy contour, the POV, and the calculated weight, all normalised to 1. Only the larger local atoms were used in this reconstruction for clarity.

We can see that the WCAD algorithm successfully decomposes the F_0 contour. The phrase component gives a qualitatively good fit to the global trend of the F_0, and the phrase-final drop is successfully captured. Also, the algorithm uses both positive and negative atoms to decompose the F_0 contour, which is in line with Strik's findings [5], and increases its physiological plausibility.

Experiment 1. The results of the first experiment are presented in Fig. 2 for the English male speaker. The figure shows plots of the $\text{WCORR}_{\text{norm}}$ relative to the number of atoms per syllable for the 10 utterances recorded by this speaker, and the average curve that is representative of how well the WCAD algorithm models this particular speaker. As expected, the WCORR curves rise steeply at the beginning as the larger local atoms are added, and eventually saturate as smaller and smaller atoms are added. It is interesting to note that saturation is reached around the 1 atom/syllable mark, which might hint at a deeper physiological plausibility of our model. Average $\text{WCORR}_{\text{norm}}$ plots were calculated for all of the speakers and are shown in Fig. 3. We can see that the average curves vary across the different speakers and languages, but that they also correlate well and follow the same general trend.

Table 2 gives the number of atoms per syllable needed on average for the WCAD algorithm to model the F_0 curve to the different perceptual WCORR thresholds presented in Table 1, for each of the speakers. Again we can see that there is a variability among the speakers, but there seems to be some correlation within the languages themselves, which is a matter for further investigation. The average atoms/syl is also given for each category.

Experiment 2. To compare the performance of our algorithm with that of the CR model, we plotted the results obtained with Mixdorff's CR parameter extraction in Fig. 3 with single points for each utterance for each of the speakers. We can see that our algorithm seems to give comparatively good results to the CR model, asserting its practical value. Note that for some

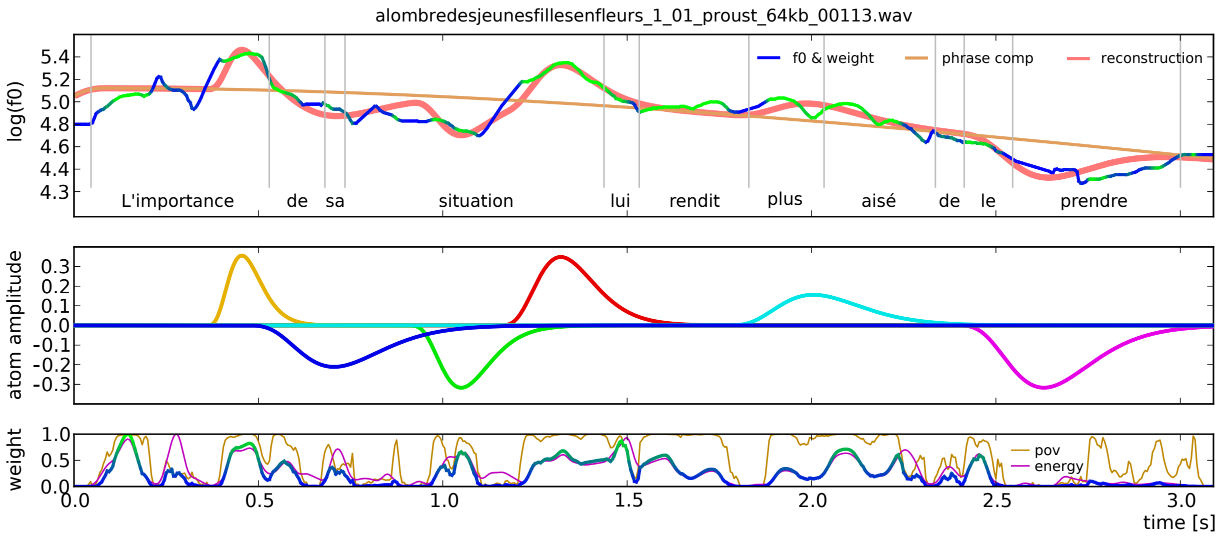

Figure 1: *Results obtained with the WCAD algorithm for the sentence* "L'importance de sa situation lui rendit plus aisé de le prendre." *by the French male speaker, showing the: original F_0, colored according to POV, phrase atom and reconstructed F_0 (top), extracted local atoms (middle), and the energy, the POV and the weighting function (bottom).*

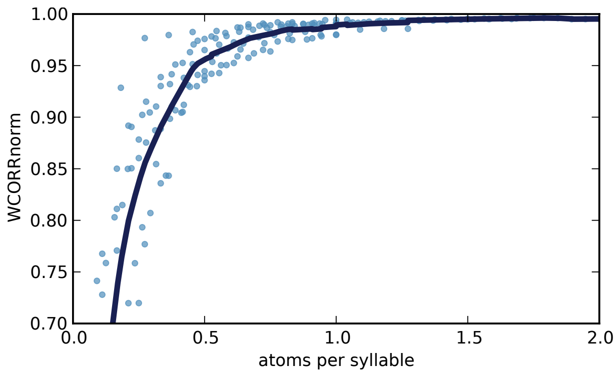

Figure 2: *Weighted correlation of the WCAD algorithm F_0 contours relative to the number of atoms per syllable for all of the sentences for the English male speaker, and the calculated average curve.*

Table 3: Average WCORR and number of atoms/syllable obtained by the CR model, for each of the speakers.

	En M	En F	Fr M	Fr F	Ge M	Ge F	Avg.
WCORR	0.973	0.964	0.967	0.976	0.967	0.969	0.969
Cat	2	2	2	2	2	2	2
commands	17	12	24	22	14	10	16
com/syl	0.46	0.42	0.37	0.37	0.42	0.47	0.42

of the utterances the extraction of the CR model parameters with Mixdorff's implementation failed and they are not included in the plot.

Table 3 gives the average WCORR and the average total number of phrase and accent commands in the CR model for each of the speakers. The results show that Mixdorff's implementation of the CR model on average gives a WCORR of 0.97, which corresponds to Category 2 from Table 1, at 0.42 atoms/syl. When comparing the average number of atoms/syl with those obtained with the WCAD algorithm, we can see that our algorithm obtains the same perceptual quality with nearly the same average number of atoms/syl.

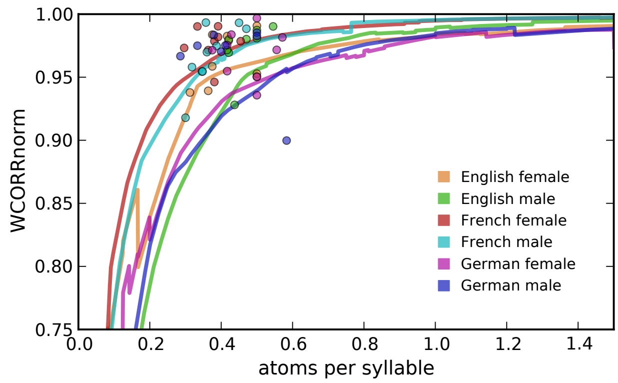

Figure 3: *Weighted correlation of the F_0 contours relative to the number of atoms per syllable averaged over all the sentences for each of the speakers in the database. The WCORRs obtained with Mixdorff's implementation of the CR model are shown for comparison.*

5. Conclusions

We have introduced a generalised CR model called the Weighted Correlation Atom Decomposition model. The model was designed to qualitatively approximate the physiological processes of intonation production. The atom decomposition process is fully automatic and is based on a matching pursuit framework, which integrates the perceptually relevant weighted correlation as a cost function. The introduced model has been shown to successfully model the intonation contours across a number of speakers and languages affirming its plausibility. In addition, it has been shown that the introduced model has comparable performance to the CR model, proving its practical value to modelling intonation in text-to-speech.

6. 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 SP2: the SCOPES Project on Speech Prosody.

7. References

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

[2] G. Kochanski, C. Shih, and H. Jing, "Quantitative measurement of prosodic strength in Mandarin," *Speech Communication*, vol. 41, no. 4, pp. 625–645, 2003.

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

[4] H. Fujisaki, "The roles of physiology, physics and mathematics in modeling prosodic features of speech," in *Speech Prosody*, Dresden, Germany, May 2006.

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

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

[7] H. Kameoka, J. Le Roux, and Y. 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.

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

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

[10] D. 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.

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

[12] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched noise data," *Hearing Research*, vol. 47, pp. 103–108, August 1990.

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

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

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

[16] S. Prom-on, Y. Xu, and B. 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.

[17] H. Mixdorff, "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, 2000, pp. 1281–1284.

[18] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[19] W. J. Hess, K. J. Kohler, and H.-G. Tillmann, "The Phondat-verbmobil speech corpus." in *EUROSPEECH*, 1995.