A NOVEL PITCH DECOMPOSITION METHOD FOR THE GENERALIZED LINEAR ALIGNMENT MODEL

Mahsa Sadat Elyasi Langarani and Esther Klabbers and Jan van Santen

Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR, USA

elyasila@ohsu.edu, klabbers@ohsu.edu, vansantj@ohsu.edu

ABSTRACT

Superpositional models of intonation typically propose decomposing fundamental frequency (F_0) contours into phrase curves and accent curves, aligned with phrases and left-headed feet, respectively. Extracting these component curves from F_0 contours without making undue assumptions is challenging. We propose a novel method for decomposing pitch curves, based on the assumption that accent curves can be described by combining skewed normal distributions and sigmoid functions. In contrast to an earlier pitch decomposition algorithm ("PRISM"), this allows for simple joint optimization of phrase and accent curve parameters, using fewer parameters. The proposed method was evaluated on three speech corpora containing: (1) synthetically generated pitch curves, (2) all-sonorant utterances, and (3) utterances containing both sonorant and non-sonorant speech sounds. The root weighted mean squared error is small, and, on the corpus for which comparable data are available, is significantly smaller than for PRISM.

Index Terms: text-to-speech synthesis, prosody modeling, superpositional model

1. INTRODUCTION

In text-to-speech synthesis, generating expressive, meaningful fundamental frequency or F_0 contours is still a challenge. For a project in our Center on generating personalized intonation using a small amount of single-speaker training data, a compact quantitative characterization of intonation contours is needed. For this purpose, we use models in the tradition of superpositional intonation models. These models posit that F_0 contours can be described as an overlay (or superposition) of component contours. Equation 1 is a formalization of this model, called the General Superpositional Model (GSM).

$$F_0(t) = \bigoplus_{c \in C} \bigoplus_{k \in c} f_{c,k}(t) \quad (1)$$

This formulation generalizes several models such as Öhman's model [1], the Fujisaki model [2, 3], and the Superposition of Functional Contours model (SFC, 4).

The Generalized Linear Alignment model (GLAM, 5, 6, 7) is a – still very general – special case of the GSM, where the $\oplus$-operator represents addition of component curves, C is the set of two curve classes representing accent and phrase curves, c is a particular curve class of C, and k is an individual curve of c's class. The phrase curve is the underlying curve that spans an entire phrase. It provides information about the baseline pitch and the global declination. The accent curves span stress-timed feet [8] and they convey the amount and type of emphasis exerted on accented syllables. A foot is defined as an accented syllable followed by zero or more unaccented syllables until the next accented syllable or a phrase boundary. In the GLAM, phrase curves are modeled as piece-wise linear (or log-linear) curves consisting of a start point, an inflection point at the start of the syllable containing the nuclear pitch accent, and an end point[9]. Accent curves are modeled as smooth single-peaked curves, but left unspecified otherwise.

The usefulness of this superpositional approach is supported by studies where the model is used to characterize prosodic characteristics of a particular speaker, emotion, or sentence [10, 11, 12], and by studies on speech synthesis to produce more natural sounding pitch contours than certain other approaches [13, 7].

Decomposing a natural F_0 contour into its component curves is challenging for two reasons. First, unless certain assumptions are made, there is no unique solution to the decomposition of a given F_0 contour because the accent curves and phrase curves can trade to produce the same F_0 contour. Second, the F_0 contour is often not smooth, interrupted by non-sonorant sounds or pauses, or perturbed by segmental effects [14].

The central goal of this paper is to describe a novel algorithm for decomposition of smoothed F_0 contours into their component curves, based on the GLAM in conjunction with relatively mild assumptions about the shapes of the underlying curves. The first implementation of a similar pitch decomposition algorithm was PRISM [15, 12, 16, 17]. It used Gaussian templates via parametrized time warp functions for modeling the accent curves. PRISM showed promising results. But, as we shall see, the specific implementation of the algorithm had some drawbacks.

In Section 2.1, we will summarize the PRISM algorithm and in Section 2.2, we will discuss our novel decomposition algorithm. Finally, in Section 3 we discuss results of using the novel decomposition algorithm on three speech corpora and perform a direct comparison between PRISM and the new algorithm on a small speech corpus containing both voiced and unvoiced segments of speech.

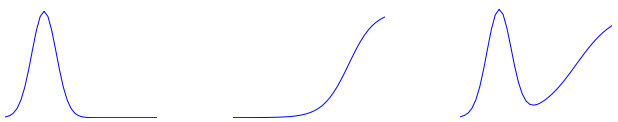
2. PITCH DECOMPOSITION

2.1. PRISM

PRISM is a pitch decomposition algorithm developed by Mishra [12]. Mishra showed that decomposed phrase and accent curves could be used in speech synthesis to create F_0 contours that sounded more natural than concatenated natural F_0 contours. However, most of the corpora on which PRISM was tested were designed not for general-purpose speech synthesis development but to exercise the method in a corpus with carefully controlled voicing and accentuation patterns. A second experiment testing perceived prominence by systematically varying GLAM parameters showed there was a significant correlation between accent curve height and perceived prominence, supporting the perceptual significance of the component curves as extracted by he method.

This material is based upon work supported by the National Science Foundation under Grant No. 0964468.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE



a: Rise-Fall b: Y/N-Question c: Continuation

Fig. 1. Three accent types in the novel decomposition algorithm

However, there are certain aspects of PRISM that must be examined more closely. First, the algorithm assumes the phrase curve is piecewise-linear, consisting of foot-length line segments instead of the two line segments allowed by GLAM. This introduces additional parameters in the process ($n + 1$ instead of 3 parameters per phrase curve containing n feet), and may also undermine the perceptual relevance of the phrase curve because there is no global declination. Second, PRISM allows negative accent curves to model F_0 values that fall under the phrase curve. American English generally does not have negative accents. Third, PRISM uses 9 parameters for estimating each accent, which, given the generally regular shapes of local pitch excursions should not be necessary – a few parameters for, e.g., location, width, and asymmetry, should suffice. Fourth, PRISM optimizes the phrase and accent curves separately, which invites local minimum problems.

2.2. The novel decomposition algorithm

The new algorithm optimizes phrase and accent curve parameters simultaneously. After initialization, the parameters for phrase and accent curves are optimized using Sequential Least Squares Programming (SLSQP, 18) which is a slight re-implementation of an algorithm proposed by Kraft [19]. There is an iterative learning process to minimize the fitting error, which is defined as the root weighted mean square error (RWMSE) as seen in Equation 2, where the weight w is computed as the multiplication of the voicing flag and the signal energy, X represents the smoothed F_0 values and Yrepresents the estimated F_0 values after decomposition.

$$RWMSE(X_i, Y_i) = \sqrt{\frac{\sum w_i (X_i - Y_i)^2}{\sum w_i}} \tag{2}$$

The process is as follows: First, pitch values are extracted using Normalized Cross-Correlation (NCC, 20) coupled with a Viterbi search to find perceptually relevant smooth pitch curves. This led to better results than using the standard get_f0 algorithm.

Second, the phrase curve parameters are initialized. Each phrase has three parameters for the start, inflection, and end point of the phrase. If the speech is voiced in those areas, the actual F_0 values are used as an initial guess. If the start of the phrase is unvoiced, the initial phrase start value is set to match the inflection point. If the end of the phrase is unvoiced, the initial phrase end value is set to match the last F_0 value in the phrase. These points are adjusted downwards if there are any F_0 values falling under the phrase curve to avoid having to model negative accent curves. The phrase curve is constructed by linear interpolation between the three points p_s, p_i, and p_e (Equation 3).

$$P(t) = interpolate(p_s, p_i, p_e) \tag{3}$$

The phrase curves are subtracted from the F_0 contour to obtain the initial values for the accent curves.

Third, we perform the curve fitting. Figure 1 shows the three accent types used: rise-fall, yes/no question, and continuation. The rise-fall accent is modeled using a skewed normal distribution with four parameters (Equation 4, A stands for the amplitude value of the accent curve, which controls the height of the curve, ξ and ω specify the location and scale, and α determines the skewness of the distribution). The question intonation is modeled using a sigmoid function with three parameters (Equation 5, B stands for the amplitude value of the sigmoid, β and γ specify the slope and location of the steepest slope) and the continuation accent is modeled as a combination of the skewed normal distribution and the sigmoid function using seven parameters (Equation 6).

$$f(t) = A\frac{2}{\omega}\phi(\frac{t-\xi}{\omega})\Phi(\alpha(\frac{t-\xi}{\omega})) \tag{4}$$

$$g(\beta(t-\gamma)) = B\,\frac{1}{1+e^{-\beta(t-\gamma)}} \tag{5}$$

$$h(t) \;=\; f(t) + g(\beta(t-\gamma)) \tag{6}$$

To initialize a rise-fall accent we compute the mean (Equation 7), the variance (Equation 8), and the skewness (Equation 9) of the F_0 values in a foot. It should be pointed out that accent curves in each phrase can overlap with each other, but the peak location of each accent curve is limited by the boundary of the foot they belong to.

$$\text{mean of } f(t):\ \xi + \omega\delta\sqrt{\frac{2}{\pi}} \quad where\ \ \delta = \frac{\alpha}{\sqrt{1+\alpha^2}} \tag{7}$$

$$\text{variance of } f(t):\ \omega^2\left(1 - \frac{2\delta^2}{\pi}\right) \tag{8}$$

$$\text{skewness of } f(t):\ \frac{4-\pi}{2}\,\frac{\left(\delta\sqrt{\frac{2}{\pi}}\right)^3}{\left(1-\frac{2\delta^2}{\pi}\right)^{3/2}} \tag{9}$$

Using Equations 1, 3, 4, 5, and 6, we can rewrite the fundamental frequency based on our function curves (Equation 10). The parameters a and b are binary to toggle between the three accent types. This equation is optimized using SLSQP.

$$F_0(t) \;= P(t) + \sum_{i=0}^{n}(b_i(a_i f(t) + (1-a_i)g(t)) + (1-b_i)h(t)) \tag{10}$$

3. EXPERIMENTS

3.1. Decomposing synthetic pitch contours

The first experiment with the novel pitch decomposition algorithm was a proof-of-concept using synthetically generated F_0 contours. The contours were generated using our text-to-speech system that uses the GLAM model to generate pitch contours. We generated synthetic curves for 229 sentences that make up the CSLU Emphasis Protocol [21]. This protocol was designed to elicit F_0 contours produced with various linguistic and prosodic features. It prescribes which syllables are accented, where each foot starts and ends, and where phrase boundaries occur. Finally, each utterance in the protocol has a target word that is spoken with a prescribed degree of emphasis. The protocol systematically varies the accent type (standard vs contrastive), the sentence type (declarative, wh-question, or yes-no question), the number of syllables in the foot (1, 2, 3 or more), and the phrasal position of the target word (initial, medial, or final). An example sentence of a wh-question with feet boundaries marked

with brackets and the target word marked in all-caps: [Will we] [really know] [MARIO], [when we're in] [Maine?].

Resynthesis with the new pitch contours as created by decomposition and superposition of the estimated phrase and accent curves showed a very small overall RWMSE of 2.5 Hz. While humans can hear very fine distinctions between two pure tones when listening to them sequentially at a short time interval, in a longer sentence this type of error is not noticeable. Klatt noted that subjects could hear a 0.3 Hz difference in a constant F_0 contour, but when the synthetic F_0 contour was a linear descending ramp (32 Hz/sec) the just-noticeable difference (jnd) slipped to 2.0 Hz [22]. Comparing perceived pitch in two sentences, it was found by 't Hart [23] that there was significant variability in the subjects' sensitivity to pitch differences. Some subjects were able to perceive differences of 1.5 - 2 semitones where others were only able to hear differences when the pitch was more than 4 semitones apart. They concluded that only differences of more than 3 semitones play a part in communicative situations. Semitones are measured on a perceptual scale and the actual frequency difference depends on the frequency range. Suppose the base frequency is 200 Hz, then a 2 semitone difference corresponds to a frequency differential of 24 Hz. But if the base frequency is really high, say 800 Hz, then the same 2 semitone differential corresponds to a frequency differential of 97 Hz.

The slight discrepancy between the generated accent curves and the decomposed curves is due to the fact that the accent curves generated by GLAM are asymmetric curves cobbled together via cosine interpolation, whereas the new decomposition algorithm uses a smooth skewed normal distribution. Not only do we suspect that this discrepancy is inaudible, we also suggest that the skewed normal distribution can provide accurate approximations to a broader range of curves.

3.2. Decomposing all-sonorant speech

The next experiment involved actual recordings using all-sonorant speech from the same CSLU Emphasis Protocol. One male speaker spoke a subset of 61 sentences in this protocol. Since the foot structure was pre-determined, labeling the feet was straightforward. The recordings were first forced-aligned to the phonemes using the CSLU Toolkit [24]. Then, the label boundaries were manually corrected and foot boundaries were inserted based on the accentuation patterns. Obviously, this manual process is in need of automation.

Figure 2 shows an example pitch contour from the CSLU Emphasis Protocol for a wh-question with rising at the end. Because the sentence has three feet, there are three pitch accents in the phrase. The dotted lines show the initial parameter estimates, and the solid lines are the final predictions. The green lines represent the phrase curve, the magenta lines represent the accent curves and the blue line represents the sum of these, forming the final predicted F_0 contour. We were only able to decompose a subset of the corpus using PRISM, so we are reporting results for that subset only (32 sentences consisting of 48 phrases). The RWMSE for decomposition using PRISM is 5.40 Hz. The RWMSE for the novel decomposition algorithm is 5.85 Hz. A two-tailed t-test revealed the difference between PRISM and the new algorithm not to be significant ($t(48) = 0.530, p = 0.597$).

3.3. Pitch decomposition of recordings with voiced and unvoiced speech sounds

In the previous experiments pitch values were available for all frames in the speech recordings, so that we could apply our optimization on continuous F_0 curves. A challenge for pitch decomposition of natural speech recordings is the presence of unvoiced regions, pauses

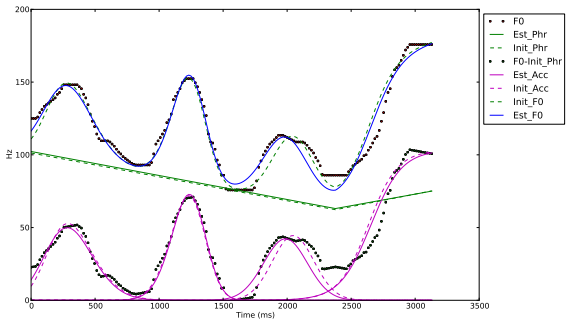


Fig. 2. Example pitch contour and decomposition for a wh-question where the pitch rises at the end: [Why will] [Molly] [run in Maine?]

where there are no pitch values, and segmental perturbations. A common way to solve this issue is to use linear interpolation between voiced areas to fill in unvoiced areas. One side-effect of having unvoiced segments in speech is that an unvoiced phoneme preceding a voiced phoneme can cause a segmental perturbation at the start of the voiced phoneme, where the observed F_0 values are slightly higher than they should be [14]. Thus, linear interpolation will give suboptimal results. Also, when a phrase starts or ends with unvoiced segments, it is challenging to find the best initialization points of the phrase curve. As it turns out, our use of weights in our RWMSE cost function substantially addresses these problems as they occur primarily in regions with low energy or voicing probability.

In order to test the novel decomposition algorithm on a speech corpus with voiced and unvoiced segments and compare it directly with PRISM, we decided to use the CSLU affect corpus which has been analyzed previously using PRISM [10]. This corpus was not specifically designed for synthesis purposes, but was created to study different affects using the same affect-neutral text for each sentence to be spoken in four different affects (Angry, Fearful, Happy, and Sad). The paper focused on one female child actor reading a total of 24 sentences in each affect (96 utterances total). The sentences are fairly short, consisting of a single phrase and 2-5 words in a phrase. The correct affect was prompted by vignettes that preceded each sentence. For this particular speaker, the pitch ranged from 200-800 Hz.

Figure 3 represents the pitch decomposition of the sentence "No way" into the component curves, for the four affect types based on the novel decomposition algorithm (Figure 3, middle row) versus PRISM (Fig. 3, top row). The bottom row represents the signal energy x voicing flag which is used as a weight on the F_0 values. PRISM detected negative accent curves for three types of affects: Happy, Fearful, and Sad. The negative accent in the last foot of the Happy sentence makes it a better fit between the actual F_0 values and the decomposed values. However, there are doubts regarding the use of negative accents in American English.

In the middle and bottom rows of Figure 3, the magenta area at the end of the F_0 curve represents a pitch rise with a weak signal amplitude. We hypothesize that this pitch rise is not perceptually relevant and as such we can ignore those pitch values in our decomposition. We added a threshold to the weight that is applied to the F_0 values to remove these less reliable pitch values. The RWMSE for our new decomposition approach and PRISM are shown in Figure 4. The novel pitch decomposition algorithm performs better than PRISM for most of the affects. The total RWMSE for the corpus is

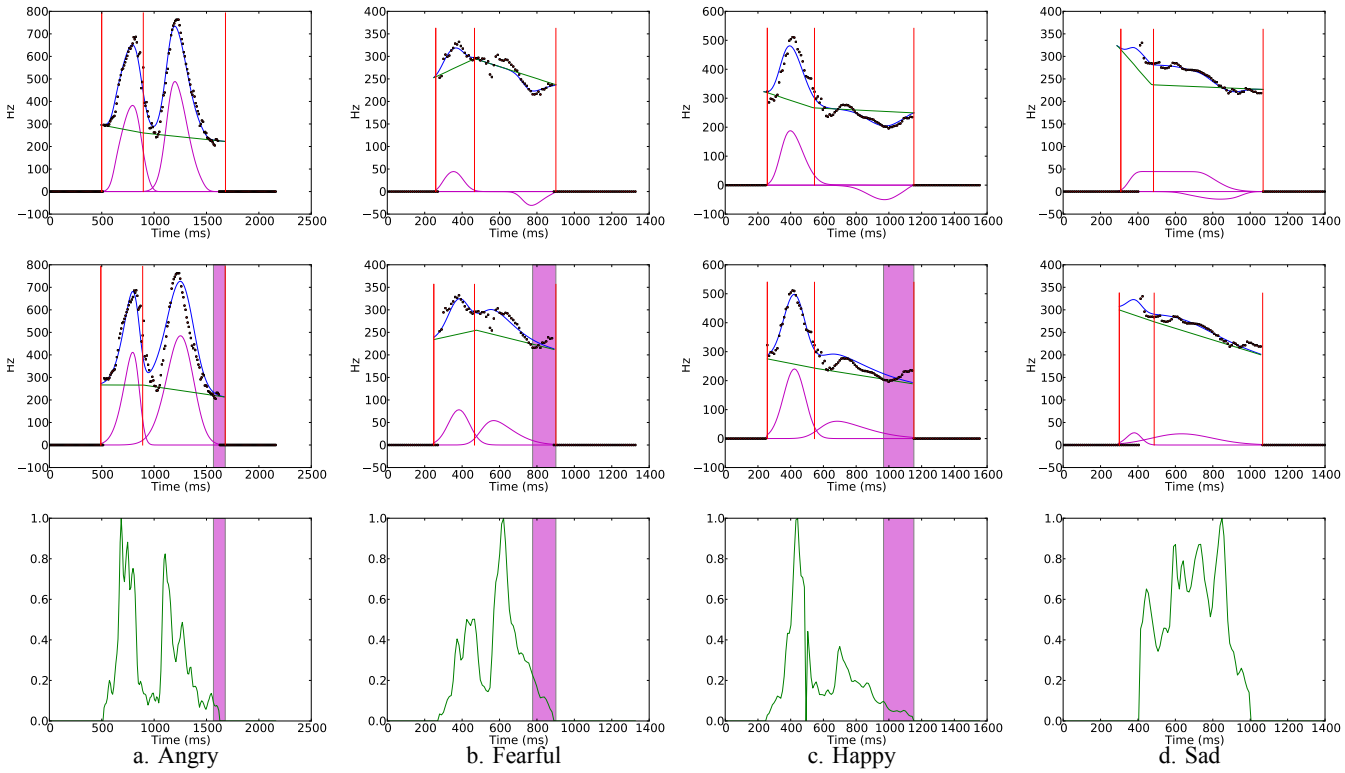


Fig. 3. Decomposition of the sentence "No way" for the four affect types (Angry, Fearful, Happy, and Sad). The first row represents the PRISM decomposition. The second row uses the novel decomposition algorithm. The third row shows the weight function (signal energy x voicing flag). The blue lines represent the estimated pitch contour, green lines represent the estimated phrase curves, magenta lines represent the estimated accent curves. The raw pitch is represented by red dots. The orange vertical lines represent the foot boundaries.

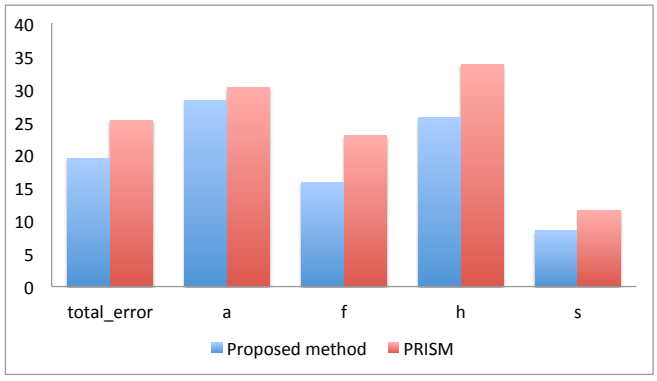


Fig. 4. The RWMSE of our method vs. PRISM. In this figure a,f,h, and s represent Angry,Fearful,Happy, and Sad groups respectively. Also the total_error shows the RWMSE of whole data

lower for our algorithm. The frequency range of the angry and fearful utterances cover the entire frequency range of the speaker (200 Hz-800 Hz) and there are a few points specifically around each accent peak that have high weight (Figure 3a) and as a result these points have more effect on the RWMSE. But since pitch perception does not follow a linear scale, pitch discrepancies at higher frequencies are likely to be less audible. Only a perceptual experiment can tell how well both pitch decomposition algorithms can reconstruct the original F_0 contour. The average difference between the RWMSE of the two methods is 5.16 Hz. We applied a one-sample two-tailed t-test to determine whether this difference is significant. The results showed that the novel decomposition algorithm performs significantly better ($t(95) = 2.22, p = 0.027$).

4. CONCLUSIONS

We proposed a novel pitch decomposition algorithm to decompose F_0 contours into phrase and accent curves in accordance with the Generalized Linear Alignment model (GLAM), a superpositional model that makes relatively mild assumptions about the shapes of the underlying component curves. We compared the algorithm with a previous estimation method, PRISM, and found that we were able to produce equivalent or better results using fewer parameters.

We are now using this method not only for projects on individualized speech synthesis for Speech Generating Devices, but also for general purpose synthesis projects based on our multi-level unit sequence approach to prosodic modeling [7].

For this method to serve as a broadly usable speech analysis tool, e.g., for emotion recognition or speaker identification, we have to demonstrate that it can be used in speech samples that are more varied and spontaneous, and less structured, than the corpora used in the present study. Usage as an analytic tool will also require automatic detection of foot boundaries, most likely requiring the integration of our method with boundary detection as well as the incorporation of other speech parameters such as energy, jitter, and shimmer.

5. REFERENCES

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

[2] H. 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.

[3] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The production of speech*, pp. 39–55, Springer, 1983.

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

[5] J. van Santen and B. Möbius, "A quantitative model of F_0 generation and alignment," in *Intonation: Analysis, Modeling, and Technology* (A. Botinis, ed.), pp. 269–288, Kluwer Academic Publishers, Netherlands, 1999.

[6] J. van Santen, C. Shih, and B. Möbius, "Intonation," in *Multilingual Text-to-Speech Synthesis: The Bell-Labs Approach* (R. Sproat, ed.), pp. 141–189, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1998.

[7] J. Van Santen, A. Kain, E. Klabbers, and T. Mishra, "Synthesis of prosody using multi-level unit sequences," *Speech Communication*, vol. 46, no. 3, pp. 365–375, 2005.

[8] D. Abercrombie, *Elements of general phonetics*, vol. 203. Edinburgh University Press Edinburgh, 1967.

[9] J. P. H. van Santen, B. Möbius, J. Venditti, and C. Shih, "Description of the bell labs intonation system," in *Proceedings of the 3th ESCA Speech Synthesis Workshop, Jenolan Caves, Australia*, pp. 293–298, November 26-29 1998.

[10] E. Klabbers, T. Mishra, and J. van Santen, "Analysis of affective speech recordings using the superpositional intonation model," in *Proc. 6th ISCA Workshop on Speech Synthesis (SSW6), Bonn, Germany*, pp. 339–344, 2007.

[11] E. Morley, J. van Santen, E. Klabbers, and A. Kain, "F0 range and peak alignment across speakers and emotions," *Proceedings of ICASSP*, 2011.

[12] T. Mishra, J. van Santen, and E. Klabbers, "Decomposition of pitch curves in the general superpositional model," in *Proceedings of Speech Prosody*, (Dresden, Germany), 2006.

[13] E. Klabbers, T. Mishra, and J. van Santen, "Recombinant speech synthesis: Natural text-to-speech synthesis with prosodic control.," *The Journal of the Acoustical Society of America*, vol. 126, p. 2205, 2009.

[14] J. v. Santen and J. Hirschberg, "Segmental effects on timing and height of pitch contours," in *Third International Conference on Spoken Language Processing*, 1994.

[15] T. Mishra, *Decomposition of Fundamental Frequency Contours in the General Superpositional Intonation Model*. PhD thesis, Oregon Health and Science University, Beaverton, OR, 2008.

[16] J. van Santen, T. Mishra, and E. Klabbers, "Estimating phrase curves in the general superpositional intonation model," in *Proceedings of the 5th ISCA Speech Synthesis Workshop*, (Pittsburgh, PA), 2004.

[17] J. van Santen, E. Klabbers, and T. Mishra, "Toward measurement of pitch alignment," *Italian Journal of Linguistics*, vol. 18, no. 1, p. 161, 2006.

[18] pyOpt developers, "pyOpt: Sequential Least Squares Programming." `http://www.pyopt.org/reference/optimizers.slsqp.html`, 2008. [Online; accessed 23-October-2013].

[19] D. Kraft, "A software package for sequential quadratic programming," tech. rep.

[20] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[21] E. Morley, E. Klabbers, J. van Santen, A. Kain, and S. Mohammadi, "Synthetic f0 can effectively convey speaker id in delexicalized speech.," in *INTERSPEECH*, 2012.

[22] D. H. Klatt, "Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception," *The Journal of the Acoustical Society of America*, vol. 53, p. 8, 1973.

[23] J. t Hart, "Differential sensitivity to pitch distance, particularly in speech," *The Journal of the Acoustical Society of America*, vol. 69, no. 3, pp. 811–821, 1981.

[24] J. Hosom, *Automatic time alignment of phonemes using acoustic-phonetic information*. PhD thesis, Oregon Graduate Institute, Beaverton, OR, 2000.