MODELING FUNDAMENTAL FREQUENCY DYNAMICS IN HYPOKINETIC DYSARTHRIA

Mahsa Sadat Elyasi Langarani and Jan van Santen

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

elyasila@ohsu.edu, vansantj@ohsu.edu

ABSTRACT

Hypokinetic dysarthria (Hd), which often accompanies Parkinson's Disease (PD), is characterized by hypernasality and by compromised phonation, prosody, and articulation. This paper proposes automated methods for detection of Hd. Whereas most such studies focus on measures of phonation, this paper focuses on prosody, specifically on fundamental frequency (F_0) dynamics. Prosody in Hd is clinically described as involving monopitch, which has been confirmed in numerous studies reporting reduced within-utterance pitch variability. We show that a new measure of F_0 dynamics, based on a superpositional pitch model that decomposes the F_0 contour into a declining phrase curve and (generally, single-peaked) accent curves, performs more accurate Hd vs. Control classification than simpler versions of the model or than conventional variability statistics.

***Index Terms*—** Hypokinetic dysarthria, Parkinson's Disease, Pitch decomposition

1. INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disease that involves the death of neurons in the substantia nigra of the midbrain, causing reduced production of dopamine which in turn results in compromised muscular coordination as well as a specific form of dysarthria, *hypokinetic dysarthria* (Hd). Features of hypokinetic dysarthria include hypernasality and compromised phonation (e.g., hoarseness), prosody (e.g., monopitch, atypical rhythm), and articulation (e.g., articulatory undershoot). Not all individuals diagnosed with PD meet criteria for Hd, although the majority will eventually develop Hd.

There is a growing interest in automatic quantitative characterization of speech in PD using signal processing and machine learning methods. Most of these studies focus on phonation, using sustained phonation recordings in which the patient is instructed to pronounce a vowel as long as possible at a steady pitch. Little et al. [1] obtained an overall accuracy of 91.4% for classification of PD vs. Control, using a kernel-support vector machine with gaussian radial basis kernel function, preceded by application of feature selection methods to a large set of features (e.g., measures of variation in fundamental frequency (F_0) and amplitude, noise to harmonics ratio, and nonlinear dynamic complexity measures). Other studies used the same data and features proposed by Little, with different classifier methods such as neural nets [2, 3], genetic algorithms [4], or decision trees [3], slightly improving on Little's results. A gaussian mixture model (GMM) based clustering approach [5] even reported 100% accuracy using 10-fold cross-validation and 50% of the data as test data.

While these results are impressive, Little's data base has four limitations. First, and most important, the overall severity level of dysarthria in the PD group is not specified. At sufficiently severe levels of dysarthria, e.g., with extreme hoarseness, the classification task not only cannot be too difficult, but also any value as a clinically useful instrument is likely to be limited. Second, the groups are not age-matched, with the ages of the PD and control groups being 67.78 and 60.25 ($t(29) = 1.94$, $p<0.06$, two-tailed) – voice changes due to aging start accelerating in the early 60's [6] ; also, the percentage of females was larger in the control group than in the PD group (62.5% vs. 30%). Third, since no additional comparison groups were included either with other types of dysarthria (e.g., spastic, flaccid, hyperkinetic, mixed) or with vocal cord dysfunction unrelated to dysarthria (e.g., polyps), the specificity of the results to PD is unknown, thus further detracting from the clinical usefulness of the methods. (The current study has the same shortcoming.) This is not to say, of course, that these methods may not be useful for *progression tracking* once a diagnosis has been established. Fourth, usage of sustained phonation recordings does not provide information about the other aspects of Hd: hypernasality, prosody, and articulation – only about phonation. The current paper focuses on prosody, specifically on F_0 dynamics.

Earlier work on F_0 dynamics has shown reduced variability in PD (e.g., [7, 8]), as expected given that monopitch is a key feature of Hd. The goal of this paper is to analyze F_0 variability in more detail, using an explicit F_0 model, called the *General Superpositional Model* (GSM). According to this model [9], the F_0 curve for a single-phrase utterance can be written as the sum of a phrase curve and any number of accent curves, one for each foot (a foot is defined as a stressed syllable followed by zero or more unstressed syllables, terminated by a phrase boundary or the next stressed syllable). This model, and its special cases, has received considerable support [9–20]. In terms of this model, reduced variability could result from atypical values of multiple components. First, a reduced slope of the phrase curve: whereas in typical speech, there generally is a declination in F_0 , perhaps the underlying factor for reduced within-utterance variability in PD is the lack of such declination. Second, reduction in the number of feet. Third, reduction in the height of either all accent curves or specific (e.g., phrase-initial, final) accent curves.

This material is based upon work supported by the National Science Foundation under Grant No. 0964468, and by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number R21DC010035 with Multi-PI's van Santen and Tjaden. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Kris Tjaden for numerous contributions, including data collection and Speech Intelligibility testing.

2. FEATURES

2.1. Baseline: Global Pitch Method

We used the per-utterance mean and standard deviation (SD) of F_0 as features.

2.2. GSM Variant #1: Modeled Accent Method

In a previous study [11], we proposed a new method to decompose the pitch contour into component curves in accordance with the GSM: a phrase curve ($P(t)$ in Equation 1) and a sum of one or more accent curves ($A(t)$ in Equation 1).

$$F_0(t) \; = P(t) + A(t) \tag{1}$$

In this method, the phrase curve consists of two linear curves, between the phrase start and the start of the phrase-final foot, and between the latter and the end point of phrase, respectively. We use the combination of skewed normal distribution and sigmoid function to model three different type of accent curves. First, the skewed normal distribution is employed to model the rise-fall accent that can happened in non-phrase-final position (f in Equation 2). Second, the sigmoid function is used to model the question intonation accent at the end of a yes/no question phrase (g in Equation 2). And, third, sum of the skewed normal distribution (f) and the sigmoid function (g) is used to model the continuation accent at the end of a non-utterance-final phrase (h in Equation 2). The number of accent curves, which is equal to the number of feet in the phrase, is shown by n in Equation 2. The values a and b are binary and can be used to compactly express the three accent types as

$$A(t) = \sum_{i=1}^{n} (b_i(a_i f(t) + (1 - a_i) g(t)) + (1 - b_i) h(t)). \tag{2}$$

For example, a yes/no question sentences with two feet (rise-fall (H*LL%) and yes/no question (L*H%) accent types) can be represented by $a_1 = 1$, $b_1 = 1$ and $a_2 = 0$, $b_2 = 1$, respectively.

$$f(t) = C \frac{2}{\omega} \phi(\frac{t - \xi}{\omega}) \Phi(\alpha(\frac{t - \xi}{\omega})) \tag{3}$$

$$g(t) = D \, \frac{1}{1 + e^{-\beta(t - \gamma)}} \tag{4}$$

In Equation 3 and 4, C and D stand for amplitude of accent curve.The two parameters set $\{\omega,\ \xi,\ \alpha\}$ and$\{\beta,\ \gamma\}$ illustrate {scale, location, skewness} of skewed normal distribution ,and {slope, location} of sigmoid function. All these parameters plus three parameters of phrase curve are optimized using Sequential Least Squares Programming (for detailed information look at [11]).

Figure 1 represents the pitch decomposition of an utterance whose content is: "All are believed to be embassy employees." into component curves, for a 49 year-old female drawn from the control group. Raw F_0 is represented by the red dotted line, and the predicted F_0 contour is represented by the blue line, which in turn is sum of the green and magenta lines, representing the phrase curve and accent curves, respectively. The gray vertical lines represent the foot boundaries. We refer to the method that uses as its features parameters extracted from the fitted accent curves (e.g., peak location; see section 3.3) as the *Modeled Accent Method*.

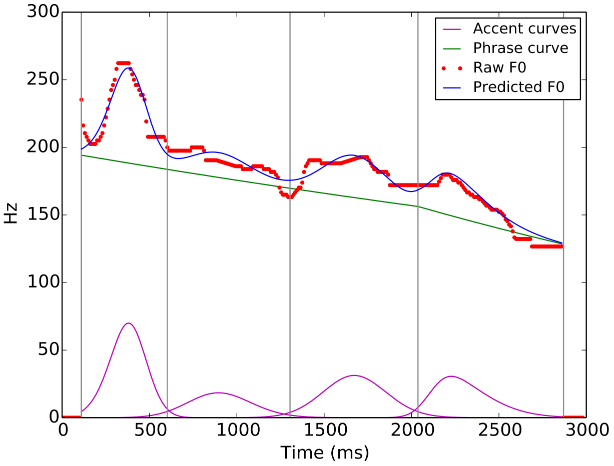


Fig. 1: Example pitch decomposition contours of a 49 year old female in the Control group for a sentence with feet boundaries marked with brackets "[All are be][lieved to be][embassy emplo][yees]."

2.3. GSM variant #2: Local Pitch Method

The *Local Pitch Method* uses a special case of the GSM, where the phrase curve is discontinuous, consisting of linear segments that each have a zero slope. In other words, the frequency value of the phrase curve in each foot is equal to the minimum F_0 value in a foot. The accent curves are obtained by, for each foot, subtracting this phrase curve from F_0. This method is used to assess the importance of a sloping, continuous phrase curve. The difference with the *Modeled Accent Method* lies in how the accent curves are computed (fitted vs. obtained by subtraction of a phrase curve from the raw F_0 curve).

2.4. GSM Variant #3: Raw Accent Method

The *Raw Accent Method* uses the same two-piece phrase curve as defined in section 2.2 in the *Modeled Accent Method*. This method is similar to the *Local Pitch Method* in that accent curves are obtained by subtraction of a phrase curve from the raw F_0 curve instead of, as in the *Modeled Accent Method*, being model based; what differs is the phrase curve shape.

3. METHOD

3.1. Participants

Participants were 10 individuals with PD (age 42-80) and 10 healthy controls (age 49-71). The average ages did not differ significantly (t(18)=1.08, p>0.25). As in the Little data, the percentage of females was larger in the control group (60% vs. 30%). Crucially, and possibly in sharp contrast to the Little data, participants were selected to have good intelligibility. And indeed, the average Speech Intelligibility values, as measured via Yorkston and Beukelman (1996)'s *Sentence Intelligibility Test* [21], were 96.3 and 97.4 in the PD and control groups, respectively (t(18)=1.21, p>0.2, two-tailed). Thus, these were groups whose speech problems, if present at all, were subtle and hence pose a challenging test for any classification algorithm (section 3.4). Using greedy text selection methods [22], we

Method	**FPos**								
	Initial				**Medial**				**Final**
	Feature	**Mean**$_{PD}$	**Mean**$_{Control}$	**P-value**	**Feature**	**Mean**$_{PD}$	**Mean**$_{Control}$	**P-value**	
Modeled Accent	loc	0.677	0.629	0.008	loc	0.435	0.462	0.080	—
					$WTSK$	0.108	0.053	0.090	
Local Pitch	$WTSD$	29.627	25.395	0.080	—				—
Raw Accent	loc	0.688	0.643	0.060	—				—
	$WTSD$	24.247	21.218	0.100					
Weighted Raw Accent	loc	0.669	0.630	0.100	$WTSK$	0.239	0.143	0.100	—
	$WTSD$	24.074	20.800	0.090					

Table 1: P-values and means for two-group, two-tailed t-tests (PD vs. Control) as a function of FPos, method, and feature; p-values larger than 0.1 are omitted.

Method	$TN(\%)$	$TP(\%)$	$Accuracy\,(\%)$	$F1$(score)
Global Pitch	30	75	52.5	0.612
Local Pitch	70	62	66.0	0.646
Raw Accent	62	61	61.5	0.613
Modeled Accent	74	69	71.5	0.708
Weighted Raw Accent	58	68	63.0	0.645

Table 2: Classification performance for each method

selected 37 sentences from the Gigaword Corpus [23] to maximally cover a (symbolic) feature space defined by features known to affect F_0 such as predicted sentence and word stress, sentence length, and word length [12]. Recordings were made in a home environment, using headsets.

3.2. Pitch tracking

We employed the Normalized Cross-Correlation method coupled with a Viterbi search to extract pitch [24]. We applied linear interpolation between voiced areas to replace the unvoiced areas.

Roughness, hoarseness, and breathiness, typical not only in Hd but also more generally in older individuals [25] increases F_0 halving and doubling [26]. Therefore, we manually corrected the extracted F_0 curves, blind as to diagnostic status (PD vs. Control). Finally we converted the F_0 values into a logarithmic scale to reduce the impact of the unequal gender distributions in the two groups.

3.3. Feature extraction

In this study, we computed features for each extracted accent curve (via the three respective GSM variants), distinguishing between feet in phrase-initial, phrase-final, and phrase-medial (i.e., neither initial nor final) position (FPos). We compute four statistical features per extracted accent curve: 1) Location (loc): location of the peak normalized by foot duration. 2) Magnitude (mag): the amplitude of the accent curve. 3) Weighted temporal standard deviation ($WTSD$): the $WTSD$ of the accent curve's distribution (Equation 5). In equation 5, t_i and x_i are the ith sample of time and accent curve value. $\bar{x}_t$ is the weighted average of time computed by ${\sum t_i x_i}/{\sum x_i}$. 4) Weighted temporal skewness ($WTSk$): the $WTSk$ of the accent curve (Equation 6).

$$WTSD = \sqrt{{\sum x_i(t_i - \bar{x_t})^2}/{\sum x_i}} \tag{5}$$

$$WTSk = \frac{{\sum x_i(t_i - \bar{x_t})^3}/{\sum x_i}}{WTSD^3} \tag{6}$$

To explore the discriminatory power of each feature, we applied t-tests to the per-speaker means of these accent features.

For SVM based classification we used larger sets of features, which also included per-speaker standard deviations (SD). $Set1$ was used for the *Global Pitch Method* and $Set2$ for the other methods, where:

$$\bullet Set_1 = \begin{cases} \text{(Per-speaker)}\; median\; of\; pitch\; mean,\; SD \\ \text{(Per-speaker)}\; SD\; of\; pitch\; mean,\; SD \end{cases}$$

$$\bullet Set_2 = \begin{cases} FPos \\ \text{(Per-speaker)}\; median\; of\; loc, mag, WTSD, WTSk \\ \text{(Per-speaker)}\; SD\; of\; loc,\; mag, WTSD, WTSk \end{cases}$$

3.4. Performance of the Global Pitch, Local Pitch, and Raw Accent methods.

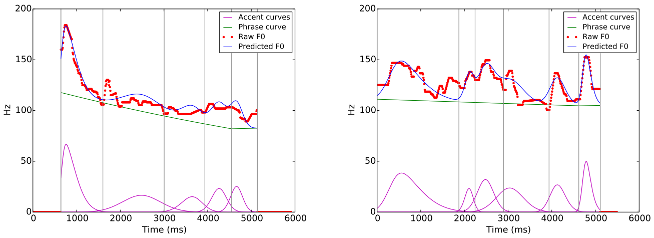
For the *Global Pitch Method*, we extracted two features, the mean and SD of the F_0 curve for each utterance (37 utterances for each speaker), and four features (loc, mag, $WTSD$, and $WTSK$) for the *Local Pitch* and *Raw Accent* methods, for each foot of each utterance (ranging from 96 to 118 feet per speaker).

We applied two-group, two-tailed t-tests (PD vs. Control) to these features. For the *Global Pitch Method* features, no significant differences were found. The third and fourth rows in the Table 1 evaluate the features derived from the *Local Pitch* and *Raw Accent* methods, and present some marginally significant results. Interestingly, only the phrase-initial feet seem to matter.

We next employed an RBF kernel based SVM using the scikit-learn toolkit [27] to classify PD vs. Control for each method. We set the $gamma$ and C SVM parameters to 10^{-1} and 10^5, respectively. We used Set_1 for the *Global Pitch Method* and Set_2 for the other methods. For evaluating the SVM results, we used accuracy and $F1$ measures. The accuracy is the average of the true positive (TP, the percentage accurate classification of participants with PD) and true negative (TN, the percentage accurate classification of control participants) rates; $F1$ is computed from Equation 7.

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{7}$$

where FP is the false positive rate $(100 - TP)$, and FN the false negative rate $(100 - TN)$. Table 2 shows for each method the averages over all selections of two held-out participants. As we expected based on the t-test results, features extracted from the *Global Pitch*



(a) Participant in Control group (b) Participant in PD group

Fig. 2: Fitted curves for two 66-year old male participants.

Method essentially yield chance performance. In contrast, features extracted from the two other methods perform better than chance. This suggests that foot-based features are more informative than global, whole-phrase features.

3.5. Performance of the Modeled Accent Method

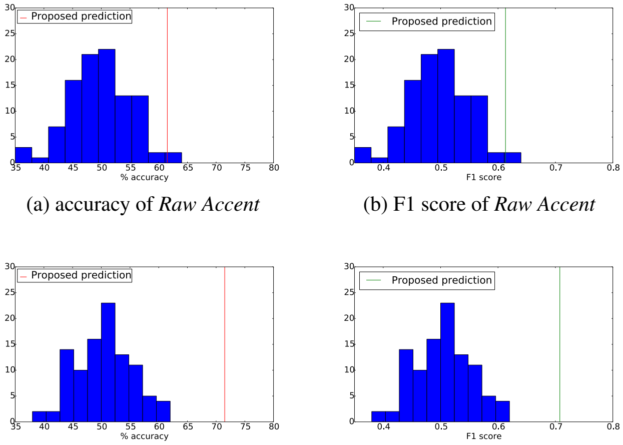
We now turn to the *Modeled Accent Method* (section 2.2). To ensure that any results are not due to a better model fit in one group, we applied a t-test to the per-participant means of the root mean square (RMS) deviation of the predicted and observed F_0 values. No significant difference between the groups was found, with the RMS values for the PD and control groups at 0.82 and 0.88, respectively. Figure 2 shows an example of pitch decomposition of a sentence "Afghan government officials were not immediately available to confirm the decision" into accent curves and phrase curve. Subfigure a and b represent the curves for two 66 year old male subjects in each group. The relatively good fit of the model is clear. We note the difference in the shape of the accent curves, specially for the phrase-initial foot.

After extracting the four standard features from the estimated accent curves (i.e., loc, mag, $WTSD$, and $WTSK$) , we applied t-tests in the same way as was done for the other models (Table 1, row labeled "Modeled Accent"). Results indicate that the groups differed significantly in the peak location of phrase-initial feet. Table 2 illustrates that the features (Set_2) extracted via the *Modeled Accent Method* yielded the highest the $F1$ score, accuracy, TN, and TP values of all methods.

For determining the significance of the classification result, we performed a randomization test in which the diagnostic status of the 20 participants was randomized 100 times and the SVM training and test procedures were applied to each randomization. Figure 3 shows the histogram of the randomized SVM results and the observed results; we display these histograms to show that the distributions resulting from randomization are well-behaved, lending credibility to this significance testing method. The histograms show that the observed results are far better than can be expected by chance for the *Modeled Accent Method*, with marginally significant results for the *Raw Accent Method*. (We used a randomization test because the assumptions underlying conventional statistical methods such as Hotelling's T^2 test are unlikely to be met.)

3.6. Improving the Raw Accent Method using Frame Weighting

In the preceding, we applied linear interpolation between voiced areas to replace the unvoiced areas. However, there may be regions that, while not fully voiceless, are nevertheless low in sonorance and thus may contribute minimally to perceived pitch and/or may include substantial segmental perturbations that are not modeled by accent curves [28]. To address this, we applied a weight to each frame obtained by the multiplication of the voiced/unvoiced flag and energy [11]. We used the same weight on the raw accent data, and then we extracted the four features (loc, mag, $WTSD$, and $WTSK$) to create *Weighted Raw Accent Method*. We applied the t-test on these four features (Table 1, last row). We found marginally significant results not only on phrase-initial feet but also on phrase-medial feet.



(a) accuracy of *Raw Accent* (b) F1 score of *Raw Accent*

(c) accuracy of *Modeled Accent* (d) F1 score of *Modeled Accent*

Fig. 3: Reliability of the classification's result

As we can see on Table 2, the results are slightly more accurate: The *Weighted Raw Accent Method* with the features (Set_2) improved the $F1$ score and accuracy 0.03 point and 1.5 precent compared to the *Raw Accent Method*. Yet, the results are still not as good as for the *Modeled Accent Method*.

4. CONCLUSIONS

The results suggest that modest levels of classification accuracy are obtained with a model based approach. Even if the accuracy is definitely too low for any practical use, the results are statistically highly significant. This is both important and surprising, given that the groups did not differ in speech intelligibility.

Importantly, the modeled accent results were better than the conventional baseline method, which uses global statistics – mean and standard deviation, or coefficient of variability. In addition, the modeled accent results were also better than those of less sophisticated methods, such as the raw accent method. Very broadly speaking, this could mean that it is in the fine details of F_0 dynamics that very mild forms of dysarthria first become visible.

Future research will focus not only on creating better speech corpora – with multiple diagnostic groups, careful age and gender matching, and much larger – but also in creating methods that address all symptoms of dysarthria: phonation, prosody, articulation, and hypernasality.

5. REFERENCES

[1] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, “Suitability of dysphonia measurements for telemonitoring of parkinson’s disease,” *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 4, pp. 1015–1022, 2009.

[2] F. Åström and R. Koker, “A parallel neural network approach to prediction of parkinson?s disease,” *Expert systems with applications*, vol. 38, no. 10, pp. 12470–12474, 2011.

[3] R. Das, “A comparison of multiple classification methods for diagnosis of parkinson disease,” *Expert Systems with Applications*, vol. 37, no. 2, pp. 1568–1572, 2010.

[4] P.-F. Guo, P. Bhattacharya, and N. Kharma, “Advances in detecting parkinson?s disease,” in *Medical Biometrics*, pp. 306–314, Springer, 2010.

[5] M. Hariharan, K. Polat, and R. Sindhu, “A new hybrid intelligent system for accurate detection of parkinson’s disease,” *Computer methods and programs in biomedicine*, vol. 113, no. 3, pp. 904–913, 2014.

[6] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, “Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age,” *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, pp. 1011–1021, 2011.

[7] G. L. Dorze, L. Ouellet, and J. Ryalls, “Intonation and speech rate in dysarthric speech,” *Journal of communication disorders*, vol. 27, no. 1, pp. 1–18, 1994.

[8] L. K. Bowen, G. L. Hands, S. Pradhan, and C. E. Stepp, “Effects of parkinson?s disease on fundamental frequency variability in running speech,” *Journal of Medical Speech-Language Pathology*, vol. 21, no. 3, pp. 235–244, 2014.

[9] J. P. van Santen and B. Möbius, “A quantitative model of fo generation and alignment,” in *Intonation*, pp. 269–288, Springer, 2000.

[10] J. P. 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.

[11] M. S. Elyasi Langarani, E. Klabbers, and J. P. van Santen, “A novel pitch decomposition method for the generalized linear alignment model,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, IEEE, 2014.

[12] J. P. van Santen and B. Möbius, “Modeling pitch accent curves,” in *Intonation: Theory, Models and Applications*, 1997.

[13] J. P. van Santen, “Quantitative modeling of pitch accent alignment,” in *Speech Prosody 2002, International Conference*, 2002.

[14] J. P. van Santen, T. Mishra, and E. Klabbers, “Estimating phrase curves in the general superpositional intonation model,” in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[15] J. P. van Santen, A. Kain, and E. Klabbers, “Synthesis by recombination of segmental and prosodic information,” in *Proceedings of the International Conference on Speech Prosody, Japan*, pp. 409–412, 2004.

[16] E. Klabbers and J. P. van Santen, “Clustering of foot-based pitch contours in expressive speech,” in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

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

[18] E. Klabbers, T. Mishra, and J. P. van Santen, “Analysis of affective speech recordings using the superpositional intonation model.,” in *SSW*, pp. 339–344, 2007.

[19] T. Mishra, J. P. van Santen, and E. Klabbers, “Decomposition of pitch curves in the general superpositional intonation model,” *Speech Prosody, Dresden, Germany*, 2006.

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

[21] K. Yorkston, D. Beukelman, and R. Tice, “Speech intelligibility test for windows,” *Lincoln (NE): Tice Technology*, 1996.

[22] J. P. van Santen and A. L. Buchsbaum, “Methods for optimal text selection.,” in *EuroSpeech*, 1997.

[23] D. Graff, J. Kong, K. Chen, and K. Maeda, “English gigaword third edition ldc2007t07,” in *Web Download. Philadelphia: Linguistic Data Consortium*, 2007.

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

[25] W. Ryan and K. Burk, “Perceptual and acoustic correlates of aging in the speech of males,” *Journal of communication disorders*, vol. 7, no. 2, pp. 181–192, 1974.

[26] A. De Cheveigné and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] J. P. van Santen and J. Hirschberg, “Segmental effects on timing and height of pitch contours.,” in *ICSLP*, 1994.