Optimizations and Fitting Procedures for the Liljencrants-Fant model for Statistical Parametric Speech Synthesis

Prasanna Kumar Muthukumar[1], *Alan W Black*[1] *H. Timothy Bunnell*[2]

[1]Language Technologies Institute, Carnegie Mellon University,Pittsburgh, USA
[2]Nemours Biomedical Research Wilmington, USA

pmuthuku@cs.cmu.edu, awb@cs.cmu.edu, bunnell@asel.udel.edu

Abstract

Every parametric speech synthesizer requires a good excitation model to produce speech that sounds natural. In this paper, we describe efforts toward building one such model using the Liljencrants-Fant (LF) model. We used the Iterative Adaptive Inverse Filtering technique to derive an initial estimate of the glottal flow derivative (GFD). Candidate pitch periods in the estimated GFD were then located and LF model parameters estimated using a gradient descent optimization algorithm. Residual energy in the GFD, after subtracting the fitted LF signal, was then modeled by a 4-term LPC model plus energy term to extend the excitation model and account for source information not captured by the LF model. The ClusterGen speech synthesizer was then trained to predict these excitation parameters from text so that the excitation model could be used for speech synthesis. ClusterGen excitation predictions were further used to reinitialize the excitation fitting process and iteratively improve the fit by including modeled voicing and segmental influences on the LF parameters. The results of all of these methods have been confirmed both using listening tests and objective metrics.

Index Terms: Speech synthesis, Liljencrants-Fant model, statistical parametric synthesis

1. Introduction

Excitation modeling is widely acknowledged to be important; yet, it has not received the amount of attention that is devoted to modelling the vocal tract. In an earlier paper[1], we were able to show that information contained in the excitation could be used to classify the emotions contained in speech. A good excitation model will therefore help us move closer towards the ultimate goal of making synthetic speech more expressive. In this paper, we attempt to construct an excitation model by using the classic Liljencrants-Fant model (LF model, for short) with a novel iterative way of estimating the parameters of this model for a given glottal flow. We begin by describing the LF model itself. We then describe a way to get a good separation between the source and the vocal tract using the Iterative Adaptive Inverse Filtering (IAIF) method[2]. The LF model is then fit to the residual obtained by the IAIF procedure. Any remnants of the residual which are not modeled by the LF model are modeled using a low order Linear Prediction fit. The vocal tract is modeled using standard LSPs[3]. The ClusterGen speech synthesizer is then used to learn a mapping between the text and the synthesis parameters. We then describe an iterative technique that we used to improve the quality of the LF model fit by initializing it with ClusterGen's predictions of the fit. We also describe optimizations that we do with ClusterGen to improve predictability.

2. Related Work

A variety of methods have been proposed and tested for estimating excitation source parameters both as a means of accurately measuring voice quality[4, 5, 6], to provide an improved source parameterization for speech synthesis[7, 8, 9, 10, 11, 12], and as a means of improving the ability to convey emotion and expressiveness in synthesis[13, 14]. In most cases, the LF model has been chosen to represent the voiced source parameters, although several alternatives have also been explored (e.g., [15, 10, 16]). The process of estimating LF parameters typically begins by inverse filtering the acoustic speech signal either manually (e.g., [8]) or automatically (e.g., [11, 12, 9]) to remove the contribution of vocal tract resonances from the signal. Thereafter, estimates of the LF parameters can be derived either directly from measures of isolated pitch epochs [11, 12], or by using search strategies that attempt to locate the best parameters by minimizing an error measure either in the time domain [4], the spectral domain [9] or a combination of the two [8].

The current approach shares features with many of these related studies, but takes a novel approach to combining voiced, mixed, and voiceless excitation by employing the LF model for a voicing source along with a low order LPC fit to information in the differentiated glottal flow that is not captured by the LF model. This approach blends smoothly from voiced to voiceless regions by allowing the LPC fit to dominate the signal in regions where the LF model parameters are difficult or impossible to estimate.

3. Liljencrants-Fant model

The Liljencrants-Fant model[17] (LF model) was developed in the 1980's as a mathematical way of describing the glottal flow *derivative*. The reason for this being that *derivative* of the glottal flow was easier to model as compared to the glottal flow itself.

The model itself consists of the following equation:

$$e(t) = \begin{cases} E_0 e^{\alpha t} sin(\omega_g t) & t < T_e \\ \frac{-E_0}{\varepsilon T_a}.[e^{-\varepsilon(t-T_e)} - \varepsilon e^{(T_c - T_e)}] & T_e < t < T_c \end{cases} \quad (1)$$

Figure 1 depicts a plot of the LF model for typical values of the parameters. The parameters T_p, T_e, T_a, T_c are explained in the figure. The parameters E_0, α and ε can be determined from the positive peak. The glottal frequency ω_g can be determined from the fundamental period T_0.

4. Iterative Adaptive Inverse Filtering

The spectra of the vocal tract and the glottal source are deeply intertwined with each other. Since the LF model was designed

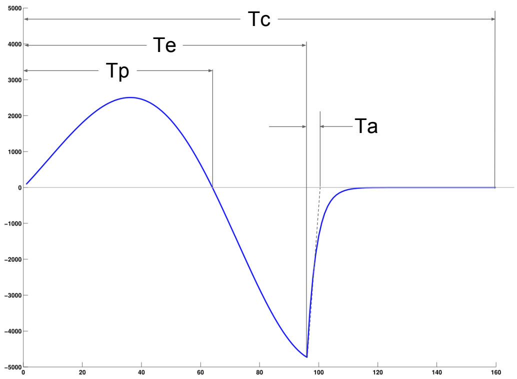

Figure 1: *LF model for typical values of the parameters*

specifically to model the glottal source derivative, we need to separate the spectrum of the glottal source from that of the vocal tract.

The Iterative Adaptive Inverse Filtering method as first described in [18] does this by iteratively making estimates of the glottal source spectrum and vocal tract spectrum and alternatively inverse filtering with one to get better estimates of the other. The final estimate of the glottal source spectrum is then far better than could otherwise have been obtained with simpler inverse filtering methods. The specific technique we use is more or less identical to the setup described comprehensively in Raitio et. al[2] but in the interests of keeping the paper as self-contained as possible, we provide a block diagram that describes the process in Figure 2.

The result of Iterative Adaptive Inverse Filtering is a set of LPC coefficients (which we convert to LSPs) that provide an estimate of the vocal tract spectrum and the glottal source function (the outputs of the two shaded blocks in Figure 2).

While it could be argued that the IAIF procedure might not produce a glottal source spectrum that is perfectly separated from vocal tract effects, we were more interested in getting a glottal source estimate that could be modeled well for speech synthesis. For that purpose, we believe that this method is sufficient.

5. Fitting LF Model parameters

Automatic extraction of LF model parameters from running speech has been approached in a variety of ways and with varying degrees of success [11]. We tried a variety of time- and frequency-domain methods for fitting LF parameters to the IAIF residual. The one reported here is the best performing for this particular task. While we found that spectral fitting worked well in some cases, it had a tendency to be quite unstable and would fail to find a proper fit in other cases. In contrast to this, one time-domain method was found to produce consistent and reasonable estimates for the LF Model parameters.

Our analysis windows were 70 msec wide and stepped in 5 msec increments through the IAIF-estimated source waveform while fitting LF model parameters. The LF parameters obtained for the single pitch pulse that spanned the center of each analysis window were output as the parameters for each 5-msec frame. Thus, the time-domain fitting, which is inherently pitch-synchronous, was used to obtain parameter estimates for a uniform frame rate. Fitting the LF model was a two-stage process in which we first located a probable instance of maximum

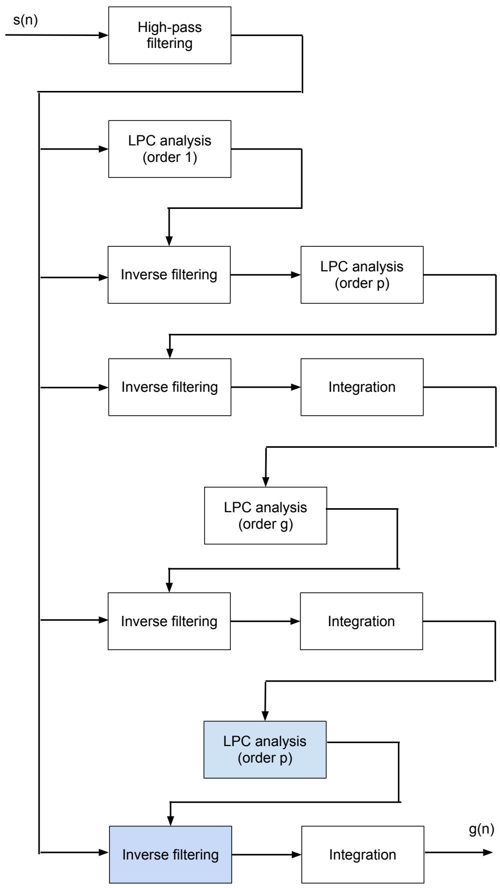

Figure 2: *Iterative Adaptive Inverse Filtering*

glottal closure rate (T_e in the LF model), and then optimized parameters for a pitch pulse around that closure event. We did this by generating a *model* LF pitch pulse with T_0 set to something close to the speaker's maximum pitch period and default shape parameters. This pitch pulse was then convolved with the windowed signal. The magnitudes, locations and spacing of peaks in the convolved signal were used to estimate voicing, pitch period locations, and F_0 within the window. If a pitch period was located in a region that spanned the center of the window, that period was then passed to a second stage optimization where the T_p, T_e, T_a, and amplitude parameters were iteratively adjusted to reduce an RMS error metric until no further improvement was found.

The second stage optimization was quite simple. The pitch period duration T_0 was held constant and the three temporal parameters, T_p, T_e, and T_a (quantized to the sampling interval for the digital waveform) were adjusted by one sample time forward or backward on each iteration. Similarly, the amplitude term was adjusted by 1dB on each iteration. A new pitch pulse was generated and the RMS error term was reevaluated following the adjustment of each parameter. If the adjustment of a parameter resulted in a reduction in the error term, the new parameter value was kept, otherwise the adjustment was discarded and the direction of change for that parameter on the next iteration was reversed. Iteration was stopped when no parameter adjustments in any direction led to further reduction in the RMS error.

This process, while very inefficient, seemed to be quite robust when supplied with reasonable starting estimates of T_0 and T_e, despite moderately strong F1-second harmonic influence on the source waveform.

We must remember that the LF model is only an *approximation* of the shapes in the glottal flow derivative. The glottal flow derivative contains a lot of high frequency content that the LF model fails to capture. Without these high frequency components, the synthesized speech tends to sound hollow and muffled. To model these components, we *subtract* the fitted LF model from the glottal derivative. A low order LPC (we used a 4th order LPC) is then fit to the remaining components that are not captured by the LF model.

Modeling it this way also has the advantage that we do not need to make a voicing decision. We had originally tried making a voicing decision and then using either the LF model or white noise depending on whether the particular phone was voiced. This approach worked really well when the voicing decisions were perfect. However, it is extremely difficult to get accurate voicing information. Whenever the voicing decision was incorrect, it either made the synthesized speech sound exceedingly jarring or sound hoarse like a smoker's voice. There is also a lot of uncertainty about what best to do in regions where the speech transitions from a voiced to unvoiced region or vice versa. These problems can be mitigated by modeling the 'residual' of the LF model.

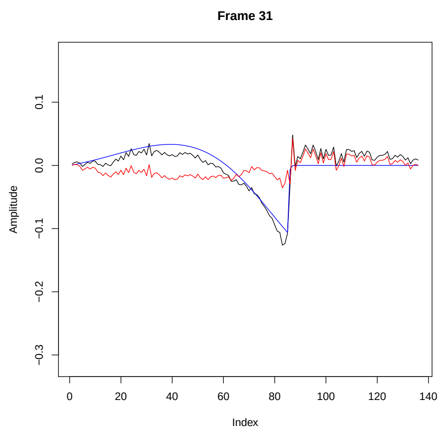

Figure 3: *Fitting to the residual: the raw glottal flow derivative is in black, the estimated LF model is in blue and the residual error is in red*

5.1. Results

To test the quality of our models, we conducted listening tests on Amazon's Mechanical Turk using the Testvox framework[19] where listeners were asked to choose between two systems. The first system was a synthesizer that used Line Spectral Pairs and the above described (LF) models for the residual. The second system was a baseline system that used an identical vocal tract model but a mixed excitation (ME) residual[20]. Listeners were asked to pick the system which they thought sounded more natural. 19 listeners were asked to choose between 10 utterances generated by each system resulting in 190 tests between the two systems. In 116 of those tests, listeners judged our system to be more natural. In 70 of those tests, listeners judged the baseline system to sound more natural. In 4 of the tests, listeners did not have a preference.

A Generalized Estimating Equation (GEE) model was used with listeners repeated over sentences to test the hypothesis that the model intercept is zero, in other words, that the odds of a listener selecting the LF version were equal to the odds of the listener selecting the ME version as more natural sounding. The model coefficient for the intercept was 0.505 with a standard error of 0.2525, a Wald Chi-Square of 4.002 with 1 degree of freedom and a p-value of 0.045. We can thus reject the hypothesis of equal odds and conclude that listeners preferred to LF model versions of the sentences.

6. Improving the Fit

The LF model fitting process described in a previous section is essentially a form of gradient descent and so is highly dependent on starting from a good initial position. Statistical Parametric Synthesis[21] provides us with an elegant way of getting good initial estimates. We start by fitting the LF model to frames of speech at 5 msec intervals, as described earlier, with a window large enough to contain at least two or three glottal pulses. We then use these parameters along with the the vocal tract model to build a synthetic voice in the ClusterGen framework. ClusterGen involves building a set of Classification And Regression Trees (CARTs)[22] that learn a mapping between the phones (with context) and the feature vectors, LSPs and LF parameters in this case.

One of the most unique parts of our approach is this: we use these CARTs to *predict* the LF parameter values of our entire database. The LF fitting process then uses these predicted parameters as seed values. These are used to create the model shape that is convolved with the frame to detect the glottal pulse and also as one of the multiple possible initializations in the gradient descent. By iteratively using CARTs to initialize the LF fitting and feeding the result of fitting process back into the CART training, we are able to get a very good final estimate of the true LF parameters. This final estimate is better than the first estimate for two reasons. Firstly, the prediction processing using the CARTs has a smoothing effect on the parameter estimates which helps to remove outliers. Secondly, the use of multiple *appropriate* initializations greatly improves the chances of the gradient descent process finding the optimum parameter values.

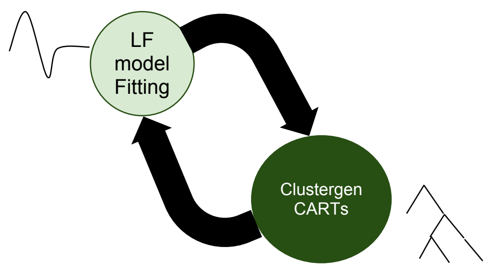

Figure 4: *Iterative estimation of LF parameters*

The ClusterGen system models the fundamental frequency F_0 independent of the other parameters. Since the LF parameters, T_0, T_e, T_p, T_a implicitly model the F_0, using these parameters directly in the ClusterGen system results in a mismatch between ClusterGen's smooth F_0 contour predictions and the implicit F_0 in the LF parameters. This causes the synthesized speech to sound shaky as if the speaker were about to cry. We can avoid this problem by representing the LF parameters as F_0-independent dimensionless parameters: Open Quotient (OQ), Speed Quotient (SQ) and Return Quotient (RQ)[23].

$$OQ = \frac{T_e + T_a}{T_0} \quad (2)$$

$$SQ = \frac{T_p}{T_e - T_p} \quad (3)$$

$$RQ = \frac{T_a}{T_0} \quad (4)$$

Using a F_0-independent representation also lets us use models like the Statistical Phrase Accent Model[24] to impose specific intonations to the speech.

6.1. Results

We iterated the fitting process several times with initializations provided by ClusterGen which were fed back to the synthesizer. Listening tests that test the naturalness of the synthesized speech were inconclusive and the difference in quality between iterations was subtle. Even speech researchers who listened to the synthesized speech from the first and last iterations acknowledged that the speech sounded different but had difficulty in making a decision on which one sounded more natural. Empirical evidence suggests that this difficulty arose from the LPC model that is used to model the remainder of the LF fitting process.

We were interested in two objective metrics. The first one was the prediction error. This is the error that we get as a result of limitations of the CARTs that ClusterGen uses. The reasoning was to try to push the fitting process into a space that ClusterGen could model well. The second metric was the RMSE and Correlation of the fitting process itself. These were computed for every pitch period where the LF model was being fit. While a low value for the second metric means the fitting process is going well, a low score for the first metric need not necessarily indicate that the fitting process works well. For example, if a bug in the code caused the parameters of every single pitch period in the database to be identical, then the ClusterGen predictions would be perfect all the time. However, we must not assume that optimizing the fitting metric alone is sufficient. The fitting process is worthless to us unless we can predict the values from the text. Therefore, our modeling technique must be able to do well under both metrics. Table 1 shows us that both the metrics indicate that our models get better as the iterations progress. Even in the case where the fitting RMSE starts to increase a little, the correlation still keeps improving.

Table 1: *RMSE and Correlation of Fitting*

	LF Fitting		Prediction
Iteration number	RMSE	Corr	RMSE
0	406.89	0.482	4.840
1	405.94	0.479	6.909
5	395.17	0.518	5.611
10	391.57	0.534	5.061
15	389.98	0.543	4.836
20	389.65	0.547	4.722
25	390.06	0.550	4.661
30	390.56	0.551	4.636

7. Optimizing the Synthesizer itself

While we use CMU's ClusterGen Statistical Parametric Speech Synthesizer [25] to model the LF parameters, there is nothing specific to ClusterGen that could not be done in any other parametric synthesizer such as HTS [26]. The only reason we used ClusterGen is because we are more familiar with our own system.

As our parameterization of speech in this model is with LSPs and LF parameters and the original phonetic and HMM state labeling is carried out with MFCCs (through EHMM [27]), we know that the labels will not be optimized for this alternative parameterization. We therefore make use of our move_label algorithm [28] which moves phoneme and HMM state boundaries based on how well the parameters at either side of the boundary can be predicted. This technique typically produces models better than the equivalent of doubling the data.

This technique has been used to optimize MCEP-based models but the LSP and LF have more varied magnitudes thus causes unintended weighting of the importance of each parameter. Thus we convert all the parameters to Z-scores (number of standard deviations from the mean), even though not all parameters are actually Gaussian. This allows a more equal optimization of the parameters. We measure LSP distortion as root mean squared difference between predicted LSPs and a held out set (in the un-Z-scored domain). We multiplied this by 1000 to give us a cosmetically nice number. For LF distortion, we again use root mean squared difference between prediction and held out data (non-Z-scored domain) but no scaling was necessary. We did 20 iteractions each.

Table 2: *Move_label metrics*

				Duration	
Optimization	Pass	LSPD	LFD	RMSE	Corr
Baseline		7.472	4.829	0.907	0.425
LSP	16	7.245	5.015	0.957	0.305
LF	20	7.961	3.702	0.961	0.310
LSP+LF	20	7.379	4.657	0.946	0.313
LSP+LF+Dur	5	7.418	4.777	0.878	0.479

Standard Deviation for the baseline system was 0.385 for LSPD and 0.589 for LFD. All other SDs are of similar magnitude

The LSP+LF, which gave the lowest LFD, produced speech which we informally identified as being more smooth than the baseline, but the durations (in text to speech) were different enough to be less desirable. When we added the duration optimization constraint to our move_label optimization it converges quicker, but the spectral quality of the signal is not perceptably different from the baseline (though the durations are better).

8. Conclusions

This work shows a novel method to derive LF parameters from appropriately extracted residuals. Addressing excitation modeling is currently one of the more important issues in statistical parametric synthesis that should offer the brightness and naturalness that high quality (and costly-to-develop) unit selection synthesizers offer. Also more importantly this will allow us to address issues in modeling different speech styles efficiently. This work also continues our direction in investigating non-standard parameterizations of speech and complementing our Statistical Phrase Accent Model for F_0 [24] where the trajectory over accents is parameterized, and articulatory feature use in statistical parametric synthesis [29].

9. References

[1] S. Steidl, T. Polzehl, H. T. Bunnell, Y. Dou, P. K. Muthukumar, D. Perry, K. Prahallad, C. Vaughn, A. W. Black, and F. Metze, “Emotion identification for evaluation of synthesized emotional speech,” in *Proc. of speech prosody*, 2012.

[2] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, “Hmm-based speech synthesis utilizing glottal inverse filtering,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 153-165, pp. 459–476, 2011.

[3] F. Itakura, “Line spectrum representation of linear predictor coefficients of speech signals,” *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[4] H. Strik and L. Boves, “On the relation between voice source parameters and prosodic features in connected speech,” *Speech Communication*, vol. 11, no. 23, pp. 167 – 174, 1992.

[5] C. Gobl and A. N. Chasaide, “Amplitude-based source parameters for measuring voice quality,” in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

[6] I. Yanushevskaya, M. Tooher, C. Gobl, and A. Ní Chasaide, “Time- and amplitude-based voice source correlates of emotional portrayals,” in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, A. Paiva, R. Prada, and R. Picard, Eds. Springer Berlin Heidelberg, 2007, vol. 4738, pp. 159–170.

[7] E. Riegelsberger and A. Krishnamurthy, “Glottal source estimation: Methods of applying the lf-model to inverse filtering,” in *ICASSP-93., 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993.*, vol. 2, 1993, pp. 542–545.

[8] J. Kane and C. Gobl, “Automatic parameterisation of the glottal waveform combining time and frequency domain measures,” *Proceedings of 6th Maveba International Workshop*, 2009.

[9] J. Kane, M. Kane, and C. Gobl, “A spectral lf model based approach to voice source paramet- erisation,” *Interspeech 2010*, 2010.

[10] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, “Utilizing glottal source pulse library for generating improved excitation signal for hmm-based speech synthesis,” in *ICASSP 2011*, 2011.

[11] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, “Towards an improved modeling of the glottal source in statistical parametric speech synthesis,” in *Proc.of the 6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007.

[12] ——, “Glottal spectral separation for parametric speech synthesis,” in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 1829–1832.

[13] Z.-H. Ling, Y. Hu, and R.-H. Wang, “A novel source analysis method by matching spectral characters of lf model with straight spectrum,” in *ACII’05 Proceedings of the First international conference on Affective Computing and Intelligent Interaction*, J. Tao, T. Tan, and R. W. Picard, Eds. Spring-Verlag, 2005, pp. 441–448.

[14] M. Tooher, I. Yanushevskaya, and C. Gobl, “Transformation of lf parameters for speech synthesis of emotion: Regression trees,” in *Speech Prosody 2008*, 2008, pp. 705–708.

[15] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, “Hmm-based speech synthesis utilizing glottal inverse filtering,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 153–165, Jan.

[16] T. Drugman and T. Dutoit, “The deterministic plus stochastic model of the residual signal and its applications,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 968–981, 2012.

[17] G. Fant, J. Liljencrants, and Q. Lin, “A four-parameter model of glottal flow,” *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[18] P. Alku, “Glottal wave analysis with pitch synchronous iterative adaptive filtering,” *Speech Communication*, vol. 19, pp. 459–476, 2.

[19] A. Parlikar. (2012) TestVox: Web-based Framework for Subjective Evaluation of Speech Synthesis. Opensource Software.

[20] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Mixed excitation for hmm-based speech synthesis,” in *Proc. Eurospeech*, vol. 1, 2001.

[21] H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[22] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and regression trees*. Chapman & Hall/CRC, 1984.

[23] G. Fant and Q. Lin, “Frequency domain interpretation and derivation of glottal flow parameters,” *STL-QPSR*, vol. 29, no. 2-3, pp. 1–21, 1988.

[24] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, “A statistical phrase/accent model for intonation modeling,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[25] A. Black, “ClusterGen: A statistical parametric synthesizer using trajectory modeling,” in *Proceedings of INTERSPEECH*, 2006, pp. 1762–1765.

[26] K. Tokuda, H. Zen, and A. Black, “An hmm-based speech synthesis system applied to english,” in *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*. IEEE, 2002, pp. 227–230.

[27] K. Prahallad, A. W. Black, and R. Mosur, “Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis,” in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[28] A. W. Black and J. Kominek, “Optimizing segment label boundaries for statistical speech synthesis,” in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3785–3788.

[29] A. W. Black, H. T. Bunnell, Y. Dou, P. Kumar Muthukumar, F. Metze, D. Perry, T. Polzehl, K. Prahallad, S. Steidl, and C. Vaughn, “Articulatory features for expressive speech synthesis,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4005–4008.