INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Semi Parametric Concatenative TTS with Instant Voice Modification Capabilities

Alexander Sorin, Slava Shechtman, Asaf Rendel[i]
IBM Research – Haifa, Israel
sorin@il.ibm.com, slava@il.ibm.com

Abstract

Recently, a glottal vocoder has been integrated in the IBM concatenative TTS system and certain configurable global voice transformations were defined in the vocoder parameter space. The vocoder analysis employs a novel robust glottal source parameter estimation strategy. The vocoder is applied to the voiced speech only, while unvoiced speech is kept unparameterized, thus contributing to the perceived naturalness of the synthesized speech.

The semi-parametric system enables independent modifications of the glottal source and vocal tract components on-the-fly by embedding the voice transformations in the synthesis process. The transformations effect ranges from slight voice altering to a complete change of the perceived speaker personality. Pitch modifications enhance these changes. At the same time, the voice transformations are simple enough to be easily controlled externally to the system. This allows the users either to fine tune the voice sound or to create instantly multiple distinct virtual voices. In both cases, the synthesis is based on a large and meticulously cleaned concatenative TTS voice with a broad phonetic coverage. In this paper we present the system and provide subjective evaluations of its voice modification capabilities.

The technology presented in this paper is implemented in IBM Watson TTS service.

Index Terms: speech synthesis, voice modification, voice transformation, glottal pulse, concatentative TTS, parametric TTS.

1. Introduction

TTS voice modification is an attractive alternative to expensive, lengthy and human labor consuming recording and processing of new speech datasets. Foreseen entertainment applications in particular will require multitudes of distinct TTS voices to be created on demand which makes the voice modification the merely viable option.

The goal of this work is to endow a product level unit selection TTS system with voice modification capabilities provided that: 1) new distinct *virtual voices* can be derived from the same TTS voice on-the-fly without the need for a target speaker data; 2) the virtual voices retain high speech quality; 3) the voice modification controls can be set externally by the user and fed to the system together with the input text. The last requirement makes the new voice creation process scalable and deployable as a part of a TTS cloud service. Such a service allows users to create new virtual voices on demand, per their imagination, needs and preferences with no need in audio recordings. The baseline product level concatenative system used in this work is the IBM TTS system [1] underlying the Watson TTS service [2].

It is known that voice modifications in general degrade the synthesized speech quality and therefore remain a research topic. Most of the voice modification research works, such as [3] and [4], deal with a *conversion* or *morphing* of a source speaker voice to a target speaker voice. This is another setup where a similarity between the modified source and target voices is the goal, and voice modifications are optimized using the target speaker audio data.

Most of the existing works that deal with the untargeted voice modification or *transformation*, such as [5], [6] and [9], explore only glottal source modifications and do not aim at a distinct voice persona creation. The latter is considered a goal in [10] but this publication reports neither the specific transformations used nor formal evaluation results.

To achieve our goal we have developed a glottal vocoder and integrated it in the baseline TTS system. The vocoder is applied to the voiced speech only, while unvoiced speech is kept unparameterized, thus contributing to the perceived naturalness of the synthesized speech. Externally configurable global transformations of the glottal source and vocal tract are applied in the vocoder parameter space.

A variety of methods has been proposed for the glottal source and vocal tract parameters estimation. Most of them rely on the Liljencrants-Fant (LF) glottal pulse model [14] and the reduced Fant model [15]. For the LF parameters estimation the works [8] - [10] employ algorithms that require precise glottal closure instance (GCI) location. The works [9] - [12] apply multidimensional optimization in the LF space while [6] and [7] use the reduced *Rd*-space [15] which enables more robust analysis procedures but is limiting in terms of the sound representation.

Aiming at an accurate and robust procedure suitable for large voice datasets processing we prefer to avoid the dependency on the GCI locations, employ only scalar optimizations but go beyond the limited *Rd*-space. To this end we have developed a novel LF parameter estimation algorithm cascading two scalar optimization procedures. At the synthesis part we propose a novel glottal pulse modification approach respecting the LF parameter constraints.

The rest of the paper is organized as follows. In section 2 we present the speech production model used. Sections 3 and 4 describe respectively the speech analysis and synthesis algorithms. The voice transformation algorithms are presented in section 5. Section 6 presents listening evaluation results.

2. Speech production model

The conventional model of voiced speech within a short time window can be represented in the frequency domain as

$$S(f) = [H(f) \cdot P(f) + N(f)] \cdot V(f) \quad (1)$$

[i] This paper is dedicated to the memory of our colleague Asaf Rendel, formerly with IBM Research – Haifa, who recently passed away.

http://dx.doi.org/10.21437/Interspeech.2017-1202

The bracketed expression represents the glottal source spectrum as a sum of the harmonic component H multiplied by the glottal pulse spectrum P and the aspiration noise component N. V is the minimum phase vocal tract transfer function. The lip radiation effect, which is approximated as the time-domain differentiation, is not explicitly present in (1) but is considered included in the glottal pulse and aspiration noise. Hereafter, referring to the glottal pulse we mean the glottal flow temporal derivative.

The harmonic component H corresponds to the time-domain delta-impulse train determined by the pitch period.

For the glottal pulse we adopt the LF model [14] illustrated on Figure 1.

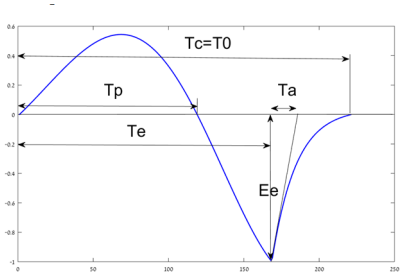

Figure 1: *LF glottal pulse derivative and its parameters.*

The LF pulse is determined by its duration which is equal to the pitch period T_0, the rise time T_p, start of the return phase T_e, the return phase duration T_a and gain E_e. Hereafter we deal with the relative dimensionless parameters $\tau_p = T_p/T_0$, $\tau_e = T_e/T_0$ and $\tau_a = T_a/T_0$ that determine the pulse shape.

In [15] Fant has developed a reduced glottal pulse space represented by a single basic shape parameter $Rd = Rd(\tau_p, \tau_e, \tau_a)$. Fant has found experimentally a reverse mapping that allows to reconstruct the LF shape parameters from the *Rd* value.

We represent the vocal tract by a high-order autoregressive model (AR) encoded by the Line Spectral Frequencies (LSF) rather than the general minimum phase model. Hence our TTS synthesizer benefits from the advantageous interpolative properties of the LSF.

3. Analysis

Pitch Estimation and Glottal Source Separation: A pitch contour is extracted from the speech signal at a 5ms-update rate using a proprietary pitch estimator [18]. All unvoiced frames are skipped. Each voiced frame is analyzed within a 3.5 pitch cycles long window.

To estimate the raw glottal signal within the analysis window we use the Iterative Adaptive Inverse Filtering (IAIF) method [16] based on the open source implementation available at COVAREP [17].

Preliminary *Rd*-pulse Fitting: The optimal *Rd*-pulse and its position relatively to the raw glottal signal are jointly estimated by maximizing the correlation coefficient between a synthetic *Rd*-pulse waveform *p(Rd)* and the pitch period long portion *g(O, T0)* of the raw glottal signal starting at a time offset *O*:

$$[Rd^*, O^*] = \arg\max_{Rd}\left[\max_{O}\frac{\langle p(Rd), g(O,T_0)\rangle}{\|p(Rd)\|\|g(O,T_0)\|}\right] \quad (2)$$

The optimization problem (2) can be solved using the simplex search method [19] implemented by the MATLAB[1] function *fminsearch*. Then a preliminary estimate of gain E_e is obtained as:

$$\widetilde{E_e} = \langle p(Rd^*), g(O^*,T_0)\rangle / \|g(O^*,T_0)\| \quad (3)$$

Aspiration Noise Modeling: The aspiration noise within a single pitch cycle is obtained by subtracting the *Rd* pulse $\widetilde{E_e}\cdot p(Rd^*)$ from the glottal signal $g(O^*,T_0)$. The noise sequence is purified by iterative outliers excluding. A 500 Hz high- pass filter is applied to the noise and the noise level is estimated as the squared energy ratio of the high-passed noise signal and the *Rd* pulse.

τ_a optimization: The *Rd*-parameterization reduces the LF space enabling tractable and robust estimation procedures. However this reduction limits the accuracy of the real glottal pulse representation. We found that the *Rd*-model captures well the general pulse shape reflected in τ_p and τ_e while in many cases it fails in representing the sharp closure phase encoded in τ_a. With fixed τ_p and τ_e the τ_a value determines the glottal pulse spectrum tilt.

Hence the next analysis step is the pulse model refinement by a τ_a optimization using the cost function in the form of the log-spectral difference between the raw and synthetic glottal signals:

$$\tau_a^* = \arg\min_{\tau_a}\|log|GS| - 0.5log[|P(\tau_a)|^2 + N^2]\| \quad (4)$$

where: GS is the raw glottal signal spectrum; $P(\tau_a)$ is the spectrum of the LF pulse with τ_p and τ_e derived from Rd^*, arbitrary τ_a and gain (3); N is the amplified ideal 500 Hz high-pass spectrum modeling the aspiration noise. The optimization (4) by the golden section search with parabolic interpolation [20] implemented by the MATLAB function *fminbnd*. An example illustrating the τ_a optimization is shown in Figure 2.

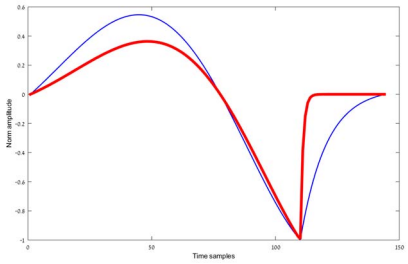

Figure 2: *A Rd-pulse (blue) and the refined LF pulse after the τ_a optimization (bold red)*

Finally, the temporal trajectories of the LF shape parameters and aspiration noise level are smoothed by a moving averaging window.

Vocal Tract modeling and source gain estimation: A power spectrum of the glottal source model is composed as the sum of the glottal pulse and the aspiration noise power spectra. A low order AR operator is derived from this power spectrum and applied to the speech signal within the analysis window. The filter output is passed to the standard LPC analysis which

[1] MATLAB is a trademark of The MathWorks, Inc. in the United States, other countries or both.

yields a high order AR operator representing the vocal tract. This AR operator is converted to a LSF vector.

The final glottal source gain E_e estimate is obtained as the squared energy ratio of the speech signal pitch cycle starting at the time offset O^* and the normalized synthetic signal. The latter is calculated using the unity gain.

The analysis yields for each voiced frame a set containing the normalized LF shape parameters, the aspiration noise level, the source gain and the vocal tract LSF vector of order 40 for 22kHz speech. This set is referred to as *Glottal Source – Vocal Tract (GSVT) frame*. The GSVT frames are stored in the voice dataset. A segment index provides a prompt access to the first and last GSVT frames associated with each speech segment.

4. Synthesis

Our semi-parametric synthesizer replaces the legacy PCM waveform generator. It receives the segment sequence produced by the TTS selection process along with duration and pitch targets. A segment is of a sub-phone level and may contain a mix of voiced and unvoiced frames. The synthesizer parses the segments and composes sequences of voiced GSVT frames and unvoiced PCM frames forming interleaving voiced and unvoiced regions. The unvoiced regions processing is inherited from the legacy PCM waveform generator. The voiced regions synthesis is explained below. The neighboring voiced and unvoiced regions are spliced together using a one pitch cycle long overlap-add window.

Global voice modification transforms described in the next section are applied to each GSVT frame.

Within a voiced region the temporal trajectories of all the modified GSVT parameters are smoothed in a vicinity of non-contiguous segment joints using a moving averaging window.

Then a voiced region is synthesized in the time-domain using an autoregressive source-filter process outlined below.

Time stamps marking pitch cycle start instants are generated according to the desired pitch curve and segment durations. A set of the GSVT parameters is calculated for each pitch cycle by interpolating respective parameter values between the frames surrounding the pitch cycle start instant. Thus the glottal source and vocal tract parameters evolve smoothly from cycle to cycle.

The aspiration noise for the entire voiced region is initialized as a 500Hz high-pass filtered Gaussian noise. The noise is amplitude modulated within each pitch cycle by the integrated glottal pulse shape as described in [7] and multiplied by a smooth gain curve derived from the cycle-wise noise levels.

A sequence of the LF glottal pulses is generated and added to the noise signal forming a glottal source component which is then amplitude modulated by a smooth curve derived from the cycle-wise gain values.

Finally the synthesized speech signal is obtained as:

$$s(t) = E(t)[p_k(t - t_k) + \sigma(t)n(t)] + \sum_m a_k(m)s(t - m) \quad (5)$$

where k is the current pitch cycle index; t_k is the current cycle start time; $p_k(t)$ and $a_k(t)$ are respectively the glottal pulse and AR operator derived from the vocal tract LSF vector; $n(t)$ is the amplitude modulated high-passed Gaussian noise; $\sigma(t)$ and $E(t)$ are respectively the noise level and gain curves.

5. Voice modifications

Pitch and duration modifications: These modifications are inherent to the GSVT synthesis process. The pitch modification is achieved by the required settings of the pitch cycle start instants. The glottal pulses uniformly stretch/shrink in time preserving their shape while the vocal tract AR operator remains unchanged. The duration modification is achieved by a proper mapping of the pitch cycles to the GSVT frames.

Glottal pulse modification: The sharpness of a glottal pulse represented in the reduced *Rd*-space can be easily modified by changing the *Rd* value as it was done in [7]. In the full LF-domain, the following constraints apply to the glottal pulse shape parameters: $2\tau_p \le \tau_e < 1 - \tau_a$

One way to assure that a modified pulse meets the constraints is to calculate it as a convex linear combination with another valid pulse in the LF space.

We set up two polar LF pulses P_{lax} and P_{tense} (shown in Figure 3) corresponding to a very lax and a very tense phonation types respectively.

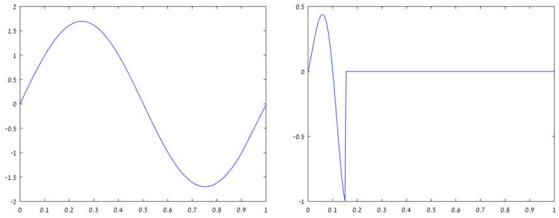

Figure 3: *Predefined polar glottal pulses* $P_{lax} = [\tau_p = 0.5, \tau_e = 0.9, \tau_a = 0.99\,]$ *on the left and* $P_{tense} = [\tau_p = 0.1, \tau_e = 0.15, \tau_a = 0.00001\,]$*on the right.*

The glottal pulse P of each GSVT frame is mixed in the normalized LF space with a polar pulse in a controlled proportion:

$$P_{mod} = \begin{cases} G \cdot P_{tense} + (1 - G) \cdot P, & if\ 0 \le G < 1 \\ -G \cdot P_{lax} + (1 + G) \cdot P, & if\ -1 < G < 0 \end{cases} \quad (6)$$

where G is an externally specified mixing factor referred to as glottal tension control. Negative/positive values of the control reduce/increase the perceived glottal tension.

Aspiration noise level modification: The aspiration noise level of each GSVT frame is multiplied by an externally specified factor referred to as breathiness control. We found that this control is particularly useful in a combination with the glottal tension modification. It allows to compensate an extra breathiness or a lack of breathiness arising from the glottal tension lowering or increasing respectively.

Vocal tract transformation: The vocal tract LSF vector of each GSVT frame is converted to the magnitude spectrum $V(f)$. A modified spectrum is calculated as $V_{mod}(f) = V(\mathrm{w}^{-1}(f))$ where $f^{out} = w(f^{inp})$ is a monotonous piece-wise linear frequency warping function passing throw predefined break points $\{(F_i^{inp}, F_i^{out} = w(F_i^{inp})), i = 1, \dots, K\}$. The break point coordinates are referred to as input and output nodes respectively. The following constraints apply to the input/output node sequence:

$$F_1 = 0 < \cdots F_i < F_{i+1} < \cdots < F_K = F_{Nyquist} \quad (7)$$

The transformed spectrum is approximated by an AR model which is converted to a modified LSF vector. The gain

of the GSVT frame is multiplied by a factor obtained as the energy ratio of the original and modified vocal tract spectra.

Informal listening reveals that 3 – 5 frequency break points with constant input nodes enable a rich repertoire of the vocal tract modification. The output nodes can be exposed for the external vocal tract modification control.

6. Subjective listening evaluation

All the evaluations reported below have been conducted using the Amazon Mechanical Turk (AMT) platform. The TTS samples have been generated from the same set of 40 sentences. The number of rates per stimuli ranged between 15 to 25 depending on the evaluation yielding from 600 to 1000 votes per system. The evaluation scores are reported along with their 95% confidence interval.

MOS scores of the semi-parametric GSVT synthesis without voice modifications compared to the baseline PCM synthesis for male and female US English voices are presented in Table 1. This evaluation was done for the research purpose only. We do not see a rational for switching to a parametric synthesis when voice modifications are not requested.

Table 1: *MOS of TTS w/o voice modification*

Male Baseline	**Male GSVT**	**Female Baseline**	**Female GSVT**
3.78±0.06	3.60±0.06	3.64±0.06	3.65±0.06

Following evaluations of the voice modification capabilities were conducted with the male voice.

The evaluation summarized in Table 2 assessed the influence of the glottal source modifications on the perceived speech quality. The combinations of down/up pitch transpositions and two-way glottal pulse modifications of equation (6) were applied to the four equal parts of the 40-sentence set.

Table 2: *MOS of TTS with glottal source modifications*

F0-25% Lax G=-0.3	**F0-25% Tense G=0.3**	**F0+35% Lax G=-0.3**	**F0+35% Tense G=0.3**
3.95±0.11	3.70±0.12	3.78±0.11	3.40±0.13

The evaluation reveals that certain glottal source modifications can improve significantly the perceived speech quality and even bring an advantage to the semi-parametric synthesis over the baseline PCM synthesizer.

Tables 3, 4 and 5 summarize an evaluation of two virtual voice examples created by a combined glottal source and vocal tract modification. The evaluation assessed the overall quality and voice distinctiveness. Both virtual voices used the glottal tension reduced with G=-0.3 (6). In a real-life application the average pitch level for a certain new voice may be constrained in order to match the (real or imaginary) visual appearance of an artificial subject. Hence the two virtual voices used a moderate pitch transposition of ±0.15% compared to the original voice. The vocal tract transformations VocTract-1 and VocTract-2 employed for the voices share the same frequency warping input nodes {200, 600, 1200, 2200, 3600} and the following output node sets respectively {250, 700, 1300, 1900, 3000} and {150, 500, 1100, 2100, 3000}.

Table 3: *MOS of TTS with two virtual voices*

Virtual voice 1: VocTract-1 Lax G=-0.3, F0+15%	**Virtual voice 2: VocTract-2 Lax G=-0.3, F0-15%**
3.51 ∓ 0.06	3.68 ∓ 0.06

Although these combined voice modifications reduced the MOS rates, the quality is still quite high in the absolute terms.

To assess the virtual voice distinctiveness a subject was presented with a pair of samples and asked to rate their voice similarity/dissimilarity on the symmetric MOS-like scale presented in Table 4.

Table 4: *Voice distinctiveness rating scale*

Definitely the same person	Probably the same person	Uncertain	Probably different persons	Definitely different persons
Rate=1	Rate=2	Rate=3	Rate=4	Rate=5

Only the verbal categories were displayed without their associated rates. The MOS-style dissimilarity scores between the baseline and virtual voices are presented in Table 5.

Table 5: *Voice distinctiveness MOS-like scores*

	Virtual voice 1	**Virtual voice 2**
GSVT orig. voice	4.29±0.08	4.33±0.08
Virtual voice 1		4.15±0.09

These examples confirm the feasibility of creation distinct virtual voices that retain high speech quality.

7. Conclusions

In this work we endowed a product level concatenative TTS system with instant voice transformation capabilities. New virtual voices can be derived from an existing large and meticulously cleaned TTS voice on-the-fly without the need for a target speaker data. The voice transformation controls can be set externally by the user. The voice transformations are applied to the voiced speech only in a glottal vocoder parameters space. The vocoder contains a novel robust glottal pulse parameter estimator. Our voice transformation repertoire includes both vocal tract and glottal source transformations using a novel glottal pulse modification algorithm. The listening evaluations confirm the feasibility of creation distinct virtual voices that retain high speech quality.

The technology presented in this paper is implemented in IBM Watson TTS service [2]. Currently three *transformable* semi-parametric GSVT voices are exposed. The transformation controls are accepted as a part of user's input composed using a Speech Synthesis Markup Language (SSML) extension.

8. Acknowledgements

We are thankful to Dr. Fernando Villavicencio who inspired us demonstrating impressive results achieved in singing synthesis based on glottal source and vocal tract parameterization and transformation.

9. References

[1] R. Fernandez, A. Rendel, B. Ramabhadran, R. Hoory, "Using Deep Bidirectional Recurrent Neural Networks for Prosodic-Target Prediction in a Unit-Selection Text-to-Speech System", in Interspeech 2015.

[2] https://www.ibm.com/watson/developercloud/text-to-speech.html

[3] Y. Agiomyrgiannakis and Z. Roupakia, "Voice morphing that improves TTS quality using an optimal dynamic frequency warping-and-weighting transform," ICASSP, 2016

[4] E. Gody, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora", IEEE Transaction on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1313–1323, 2012.

[5] S. Huber, A. Roebel, "On glottal source shape parameter transformation using a novel deterministic and stochastic speech analysis and synthesis system", INTERSPEECH 2015

[6] G. Degottex, A. Roebel, X. Rodet, "Phase Minimization for Glottal Model Estimation", IEEE Transactions on Audio, Speech, and Language Processing 19, 1080–1090, 2011.

[7] G. Degottex, P. Lanchantin, A. Roebel, X. Rodet, "Mixed source model and its adapted vocal-tract filter estimate for voice transformation and synthesis", Speech Communication, vol. 55, no. 2, pp. 278–294, 2013

[8] J. Cabral, K. Richmond, J. Yamagishi, and S. Renals, "Glottal spectral separation for speech synthesis," IEEE Journal of Selected Topics in Signal Processing, vol. 8, pp. 195–208, 2014.

[9] D. Vincent, O. Rosec, T. Chonavel, "A new method for speech synthesis and transformation base on an ARX-LF source-filter decomposition and HNM modeling", In Proc. ICASSP 2007.

[10] Y. Agiomyrgiannakis, O. Rosec, "ARX-LF-based source-filter methods for voice modification and transformation", In Proc. ICASSP 2009.

[11] T. Raitio et al, "HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering", IEEE Trans on Audio, Speech, and Lang. Proc, vol. 19, no. 1, Jan. 2011.

[12] P. K. Muthukumar, A. W. Black, T. Bunnell, " Optimizations and Fitting Procedures for the Liljencrants-Fant model for Statistical Parametric Speech Synthesis", In Proc. INTERSPEECH 2013.

[13] L. Juvela, B. Bollepalli, M. Airaksinen, P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using deep neural network", In Proc. ICASSP 2016.

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

[15] G. Fant, "The lf-model revisited. transformations and frequency domain analysis," STL-QPSR Journal, vol. 36, no. 2-3, pp. 119–156, 1995.

[16] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," Speech Communication, vol. 11, pp. 109–118, 1992.

[17] COVAREP: A Cooperative Voice Analysis Repository for Speech Technologies http://covarep.github.io/covarep

[18] D. Chazan, M. Zibulski, R. Hoory, G. Cohen, "Efficient periodicity extraction based on sine-wave representation and its application to pitch determination of speech signals", in Proc. Eurospeech 2001.

[19] J. Lagarias, J. Reeds, M. Wright, P. Wright, "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions", SIAM Journal of Optimization, Vol. 9 Number 1, pp. 112-147, 1998.

[20] R. Brent, "Algorithms for Minimization without Derivatives", Prentice-Hall, Englewood Cliffs, New Jersey, 1973