Parametric Analysis of Speech Signals Based on Estimation of Joint Source-Filter Model Using Evolutionary Computation

Mário Uliani Neto, José Eduardo de C. Silva, Diego A. Silva,
Leandro de C. T. Gomes, Thiago de A. M. Campolina
CPqD Foundation
Campinas, SP, Brazil
Email: {uliani, jcsilva, diegoa, tgomes, thiagomc}@cpqd.com.br

Hani C. Yehia*, Maurílio N. Vieira*, João P. H. Sansão†
**DEE – UFMG*
† DETEM – UFSJ
Email: {hani, maurilionunesv}@cpdee.ufmg.br
joao@ufsj.edu.br

***Abstract*—This paper presents an analysis-by-synthesis algorithm based on joint estimation of speech parameters. The main advantage of the proposed algorithm is that both vocal tract and glottal source parameters are estimated simultaneously in an automatic way. The use of evolutionary algorithms is proposed to optimize these parameters. The results show that this strategy seems to be feasible for some applications such as compression of speech signals and voice conversion.**

***Keywords*-*Analysis-by-synthesis*; *evolutionary algorithm*; *speech signal processing*; *source-filter model*.**

I. Introduction

One of the approaches used to model the process of speech production is the so-called source-filter model [2]. In this model, the human vocal tract is separated into two distinct components: a linear filter, whose transfer function is related to the resonance frequencies of the supra-glottal cavities in the human vocal tract (mouth, throat, nasal tract), and a generating source that excites the filter with an input signal.

The type of signal emitted by the source depends on the characteristics of the speech signal to be analyzed. In voiced speech segments, whose typical example is vowels, the source signal is almost periodic, due to the vocal fold vibration. In unvoiced segments, such as the fricative consonants /s/ and /f/, the signal is treated as a white Gaussian noise. For hybrid segments, the source signal is seen as a sum of the two components described above.

On the basis of the source-filter model paradigm, speech signals are analyzed through the estimation of parameters for the excitation source and the vocal tract filter, by minimizing an error measure between the original signal and the one produced when the source signal is applied to the filter.

In this study, the representation of the voiced portions of speech signals is performed by means of a simplified source-filter model, proposing the use of evolutionary algorithms to jointly estimate the parameters of the source and the vocal tract filter [12]. The model is based on physical characteristics of the speaker: the vocal tract model is able to identify the formants of the speech signal, while the excitation model makes use of the glottal waveform, as well as of the aspiration and frication noises. To best fit the latter, the use of the Transient Modeling Synthesis (TMS) algorithm is proposed.

The paper is structured as follows. Section II presents a state of the art. Section III introduces the proposed analysis-by-synthesis algorithm. In Section IV, experimental results are presented to illustrate the proposed method. Finally, Section V presents conclusions and perspectives for future study.

II. State of the Art in Joint Source-Filter Estimation

Among the approaches used in joint source-filter estimation, the most common is based on the use of models for estimating the glottal waveform and the vocal tract, defining an error function and describing techniques for optimizing it. Next, some works addressing these points are presented.

In [3], the vocal tract is modeled by a filter with poles and zeros and the glottal waveform is obtained by the LF model [2]. All the parameters are estimated minimizing the least square error (LSE).

In [5], the glottal model is approximated by the Rosenberg-Klatt model (RK) [1], and the vocal tract is modeled by a Kalman filter. The Simulated Anealing algorithm was used to find the best set of parameters.

Lu and Smith [6] proposed a convex optimization method for estimating the parameters of the source and the vocal tract jointly. They used the RK model to estimate the glottal signal, and an all pole filter to vocal tract. The error criterion is the difference between estimated and original waveforms. Del Pozo and Young [10] use a similar method, but suggested certain improvements int the the glottal source model.

The method proposed in this paper presents some advantages, and the main ones are:

- The proposed vocal tract model automatically identify the resonance frequencies of the vocal tract, unlike the techniques presented. These frequencies carry information of the physical structure of the speaker's vocal tract. This information is important in voice morphing applications.
- In some studies (i. e., [6], [10]), the parameter that indicates the time of the glottal closure, GCI (glottal closure instant), is estimated a priori. The proposed

ISBN: 978-1-61208-249-3

joint estimation method, based on evolutionary algorithms, enables the optimization of GCI together with the source and vocal tract parameters.

- The estimation method proposed in this paper, based on evolutionary algorithms, incorporates the spectral tilt coefficient in the joint optimization. Thus, this parameter can change over the time. In other papers, this parameter is usually a constant.

III. Analysis by Synthesis Based on Joint Source-Filter Estimation

The speech production model proposed in this article is illustrated on Figure 1. The voiced and unvoiced portions of the speech signal are modeled in different ways. For the voiced segments, the derivative of the glottal waveform is modeled by means of the Liljencrants-Fant (LF) model [2]. The aspiration and frication noises are modeled using TMS and a white Gaussian noise with modulated amplitude. The vocal tract is modeled as a filter containing only poles, composed of two structures: one based on formant frequency and bandwidth, and an additional one representing the information not covered by the formant filter. For unvoiced frames, the turbulence noise is modeled as a white Gaussian noise, while the vocal tract is modeled as a filter containing only poles. The details of the proposed system are presented in the following subsections.

A. Joint Source-Filter Deconvolution Based on Evolutionary Algorithms

The method developed to estimate the parameters of the glottal source and vocal tract are based on a paradigm known as source-filter deconvolution. The proposed deconvolution algorithm is based on evolutionary computation [9]. The filter parameters of the vocal tract (modeled by a set of formants in cascade), the parameters of the excitation source (in this step, for simplicity, the Rosenberg-Klatt (RK) model is used to model the source) and the GCI are jointly estimated.

1) Glottal Source – RK Model: The RK model, parameterized in the time domain, models one phonatory period of the derivative of the glottal waveform, accounting for the glottal opening and closure instants. The RK model is described by 4 parameters: a, T_0 (the phonatory period), n_c (duration of the closed phase) and μ (control the spectral tilt).

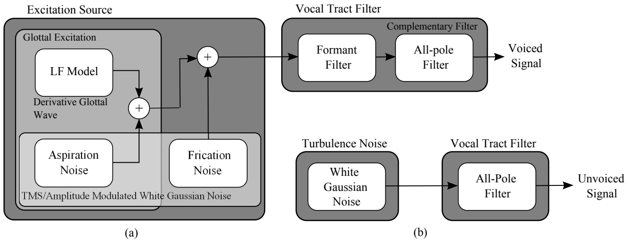


Figure 1. Source-filter model. a) Model for hybrid and voiced signals. b) Model for unvoiced signals.

Table I
Source and GCI parameters.

Parameter	Restriction	Enconding
n_c	$0.4 \times T_0 < n_c \leq 0.7 \times T_0$	Integer
a	$a \leq 0$	Real
μ	$0 < \mu < 0.99$	Real
GCI	$0 \leq GCI \leq T_0$	Integer

Table II
Restrictions on frequency and bandwidth for the cascade formant filter.

Formants	Freq (Hz)		BW (Hz)	
	Min	Max	Min	Max
1	150	900	40	500
2	500	2,500	40	500
3	1,300	3,500	40	500
4	2,500	4,500	100	500
5	3,500	5,500	150	700

2) Vocal Tract Model – Formant Filter: The vocal tract filter consists of a set of resonators in cascade [1]. Each resonator is specified by two parameters: the resonance (formant) frequency F and resonance bandwidth BW.

3) Optimization Based on Evolutionary Algorithms: The method for the joint estimation of the glottal source and vocal tract filter parameters is based on evolutionary computation. To reduce the number of parameters to be optimized and limit their search space, reducing computational complexity, simplified models are used for the excitation source (Section III-A1) and the vocal tract filter (Section III-A2). A set of restrictions is also assumed, so that the models represent a valid physical structure; this provides an additional gain in computational cost and decreases the amount of local minima in the fitness surface. The RK model is used in this step with the restrictions shown in Table I.

The vocal tract is modeled through a filter with the restrictions listed in Table II [1].

An evolution strategy (μ + λ) was used to simultaneously optimize model parameters. Both the source parameters (n_c, a, μ, GCI) and the filter (F_i, BW_i) were part of the chromosome, so that the models could be jointly evaluated.

The recombination pairs were selected based on a uniform probability. Thus, the probability of choosing an individual was the same for all the population. A discrete recombination operator, which generates the child genetic material selecting the genes of parents with equal probability, was used.

The chosen mutation operator acts on each gene of the entire population. The mutation consists of applying a gaussian mutation in all the parameters of the entire population of chromosomes. Each gene is mutated according to its variance, which is stored in the chromosome. After all the chromosomes of the population mutates, it is checked if they are still feasible. If one or more chromosomes become unfeasible, they have the value of its mutated gene restored to the value used before the mutation.

The purpose of the evolutionary algorithm is to minimize the error between the original and the estimated signal. The equation for calculating the fitness of a solution is presented in equation 1. This function was constructed to obtain higher fitness values for solutions that present lower squared error between their samples and the original samples.

$$fitness = \frac{1}{1 + \sqrt{\sum_{i=1}^{n} (s(i) - \hat{s}(i))^2}}, \qquad (1)$$

where $s(i)$ represents the original speech frame and $\hat{s}(i)$ the estimated speech frame.

More details of this implementation were presented in [11].

B. Adjustment of the Vocal Tract Using Adaptive Filters

To improve the vocal tract model, we propose the use of an adaptive filter that can estimate the parameters of an additional filter, which models the characteristics of the vocal tract that are not present in the formant filter. The adaptive filter is based on the Wiener filter, optimized with the Recursive Least Squares (RLS) algorithm. Figure 2 details this filtering model.

This method is based on the comparison between the output $g_o(n)$ of the Wiener filter and the estimate of the glottal source derivative $\hat{g}_{RK}(n)$, obtained in the process of joint estimation. When the Wiener filter coefficients are properly adjusted, the error signal $e(n)$ is minimized, which implies a filter output signal $g_o(n)$ as close as possible to the signal $\hat{g}_{RK}(n)$. The signal $g_o(n)$ is the derivative of the glottal waveform, obtained through the deconvolution of the original speech frame with the vocal tract filter (composed of the formant filter and the additional filter).

Due to inaccuracies caused by the simplification of the formant model used for the vocal tract, the joint estimation process does not produce a good estimate of the interval during which the glottis remains closed (parameter n_c).

To adjust n_c, we propose a linear search process during filter adaptation. The adaptive filtering algorithm is executed for values of n_c between 40% and 70% of the phonatory period [4], with step of 10%. The optimal solution is the one that leads to the lowest mean square error between the original frame and the temporal waveform of the signal synthesized from $\hat{g}_{RK}(n)$ and the vocal tract filter.

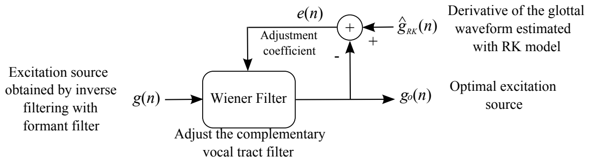

$e(n)$
+
$\hat{g}_{RK}(n)$
Derivative of the glottal waveform estimated with RK model
Adjustment coefficient
-
+
Excitation source obtained by inverse filtering with formant filter
$g(n)$
Wiener Filter
$g_o(n)$
Optimal excitation source
Adjust the complementary vocal tract filter

Figure 2. Adaptive filter used to adjust the additional filter in the vocal tract model.

C. Adjustment of the Glottal Source - LF Model

The LF model [2] is able to describe the glottal waveform derivative with greater precision than the RK model.

The waveform produced by the LF model can be determined from the four temporal parameters $\{T_p, T_e, T_a, T_c\}$ and the magnitude of $\{E_e\}$.

The fit of the LF model is performed in two stages. Initially, the LF model parameters (T_p, T_e, T_a, T_c, E_e) are estimated through direct methods [8]. This technique measures the parameters directly from the waveform obtained by means of the RK model.

The LF model parameters are then refined through an evolution strategy. The estimated T_c and T_e are considered reliable [8] and remain unchanged in the optimization process. The parameter T_a is confined between 0 and $T_c - T_e$; T_p can vary within $\pm 20\%$ of its initial estimate; and E_e can vary within $\pm 10\%$ of its initial estimate. The fitness function is based on the quadratic error between the derivative of the glottal waveform of the original frame and the waveform of the model adjusted through the LF model.

D. Modeling of the Residual Noise

The modeling techniques for the glottal source described in the previous sections do not account for the aspiration and frication noises. Due to this limitation, the difference between the source signal $\hat{g}_{LF}$ and the signal obtained through the inverse filtering of the original frame produces a residual noise. One way to interpret this noise is to consider it as a white Gaussian signal modulated in amplitude [10]; however, this approach does not always produce satisfactory results. In order to handle the residual noise, the use of the TMS algorithm [7] is proposed.

This technique exploits the time-frequency duality of sinusoids and impulses, and, as shown in [7], can be used to separate segments showing impulsive characteristics in the time domain. This process is illustrated in Figure III-D.

E. Modeling the Final Residue

The final residue model used in this article is based on the technique proposed in [10]. The method consists in treating the residue as a Gaussian noise synchronized with the phonatory period and modulated in amplitude by the LF model.

The final residue is parameterized as follows: first, a Gaussian noise with zero mean and unit variance is modulated by the glottal waveform obtained from the LF model. Then, the energy of this modulated noise is set to be equal to the energy of the final residue.

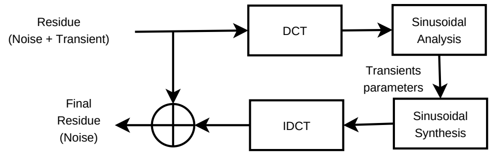


Figure 3. Block diagram of the TMS operation.

IV. Results and Analysis

For the evaluation of the proposed model, we present results for two signals: the first is a synthesized signal sampled at 8 kHz, generated by an LF source with a frequency of 120 Hz and a tract with four formants located at 500 Hz, 1.500 Hz, 2.500 Hz, and 3.500 Hz, with a bandwidth of 100 Hz; the second one is a frame of a stressed vowel /a/, extracted from a studio-quality recording of an utterance produced by a male speaker, encoded in linear PCM at a sampling rate of 8 kHz with 16 bits per sample.

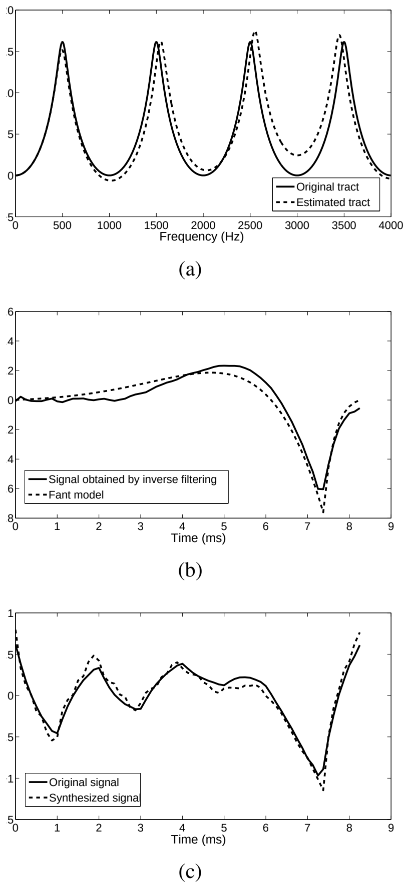
For optimization, we set an evolution strategy to iterate through 100 generations, assuming a total of 200 individuals (μ) per generation. In each generation, 400 children (λ) were created, with a crossover rate of 1 and a mutation probability of 1 (these values were empirically adjusted).

Figure 4 shows the results for the synthesized signal. As can be seen in 4(a), the estimated tract is very close to the original one. There are small differences that are especially concentrated in high-order formants, which may be justified by the fact that the fitness of the evolutionary algorithm (section III-C) is based only on the square error of the glottal signal. Most of the glottal pulse energy is concentrated at low frequencies ($f < 2.000$ Hz); thus, an error in high order formants does not significantly affect the glottal signal, which in turn has no expressive effect on the fitness. Figures 4(b) and 4(c) show that the estimate of the derivative of the glottal waveform and the synthesized signal are close to their references. The differences may be explained by imprecisions in the optimization of the tract that reflect on the glottal signal. Nevertheless, the synthesized signal is close to the original one, as shown in Figure 4(c).

Figure 5 represents the optimization of a signal with source and tract identical to the previous ones (figure 4), except for the addition of aspiration noise to the glottal source. The aspiration noise was generated as a Gaussian noise of zero mean and unit variance, modulated by the glottal signal. Figure 5(a) displays the original time signal and the signal recovered by the algorithm (including the use of the TMS algorithm). It can be observed that, though the signals differ, they present similar contours.

Figure 5(b) shows the autocorrelation function of the noise in the signal (calculated as the difference between the LF model and the signal obtained by inverse filtering the original signal) with and without TMS. Unlike the autocorrelation function of the noise produced without TMS, the autocorrelation of the noise generated with TMS remains within the range of 95% (defined by the horizontal dotted lines present in the figure) for all delays greater than two samples, which means that the generated noise is whiter when TMS is used. It is important to state that the evaluation of the impact of TMS parameters on the presented results is not the focus of this study.

Figure 6 presents a preliminary result of optimization for a frame of a real signal of the vowel /a/. The waveform of the synthesized signal presents, notably at the beginning and end of the figure, large deviations from the original signal. This probably happened because of limitations in the models used in this work. The precise identification of the reasons for this mismatch is a subject of future research.



(a)

(b)

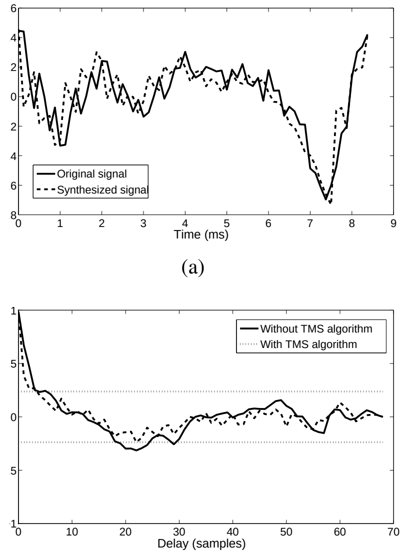
(c)

Figure 4. Optimization of a synthesized signal. a) Vocal tracts. b) Glottal signals. c) Temporal waveforms.

V. Conclusions

This article presented a technique for the joint optimization of the sound source and vocal tract filter parameters for the production of voiced portions of speech signals. The LF model was employed to represent the source, while a strategy based on the TMS algorithm was used for modeling noise components. The filter was obtained by cascading four formants, defined by their bandwidths and center frequencies. The optimization was performed through an evolution strategy.

Evolutionary computation has the advantage of allowing the joint optimization of source and filter parameters, even if the fitness function implies a multimodal problem. In all simulations, the algorithms found solutions that were feasible and satisfying. The main disadvantage of this approach is related to computational cost for real time applications, requiring a significant amount of time for the convergence of the algorithms.



(a)

(b)

Figure 5. a) Original signal and recovered signal using the TMS algorithm. b) Autocorrelation function of the noise signal.

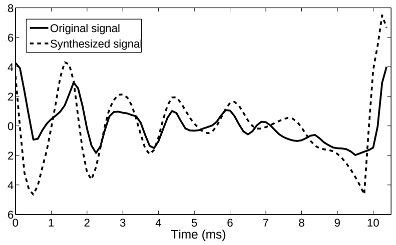


Figure 6. Temporal waveforms of the real and the synthesized signals (one frame of the vowel /a/).

One advantage of the optimization method presented in this article is its ability to efficiently determine the optimal spectral decay coefficient, as well as the instant of glottal closure; in other studies, as presented in [10], it is common to use specific algorithms for the obtainment of these parameters. Furthermore, the use of the TMS algorithm enables the whitening of the residual noise obtained as the difference between the derivative of the original glottal waveform and the one produced by the LF model, leading to a better adjustment of the noise.

The approach presented in this article for modeling the source and the filter allows the physical interpretation of the parameters obtained from the optimization, since the LF model expresses the derivative of the glottal pulse and the cascade formant filter represents the spectral envelope of speech frames. This technique seems to be feasible for applications such as compression of speech signals, voice conversion (as the parameters obtained from the optimization can be modified), and smoothing of frame boundaries prior to concatenation (for concatenative speech synthesis).

In the next steps of this research, the feasibility of applying the proposed technique in voice conversion systems will be analyzed.

REFERENCES

[1] H. Klatt, *Software for a cascade/parallel formant synthesizer*. Journal of the Acoustical Society of America, volume 67, issue 3, pp. 971-995, March 1980.

[2] G. Fant, J. Liljencrants, and Q. Lin, *A four-parameter model of glottal flow*. STL-QPSR, volume 26, issue 4, pp. 1-13, 1985.

[3] A. K. Krishnamurthy, *Glottal source estimation using a sum-of-exponentials model*. IEEE Transactions on Signal Processing, volume 40, issue 3, pp. 682-686, Mar 1992.

[4] M. N. Vieira, *Automated measures of dysphonias and the phonatory effects of asymmetries in the posterior larynx*. Ph.D. dissertation, University of Edinburgh, Scotland, 1997.

[5] W. Ding, N. Campbell, N. Higuchi, and H. Kasuya, *Fast and robust joint estimation of vocal tract and voice source parameters*. In IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-97, pp. 1291-1294, Munich , Germany, Apr 1997.

[6] H. L. Lu and J. O. Smith, *Joint estimation of vocal tract filter and glottal source waveform via convex optimization*. In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 79-82, New Paltz, NY, 1999.

[7] T. S. Verma and T. H. Y. Meng, *Extending spectral modeling synthesis with transient modeling synthesis*. Computer Music Journal, volume 24, issue 2, MIT Press, pp. 47-59, 2000.

[8] J. Perez and A. Bonafonte, *Automatic voice-source parameterization of natural speech*. In Proceedings of Interspeech, pp. 1065-1068, Lisbon, Portugal, Sep 2005.

[9] L. N. de Castro, *Fundamentals of Natural Computing: Basic Concepts, Algorithms and Applications*. Florida, USA. Chapman & Hall/CRC, 2006.

[10] A. Del Pozo and S. J. Young, *The linear transformation of LF glottal waveforms for voice conversion*. In Proceedings Interspeech, pp. 1457-1460, Brisbane, Australia, Sep 2008.

[11] M. Uliani Neto, B. Costa, F. Simões, R. Violato, and M. Leal *Estimação conjunta do processo de produção de sinais de fala utilizando computação evolutiva*. In IV Congresso Tecnológico InfoBrasil, Fortaleza, Brazil, 2011.

[12] M. Uliani Neto, José E. de C. Silva, Leandro de C. T. Gomes, Diego A. Silva, Thiago de A. M. Campolina, João P. H. Sansão, Hani C. Yehia and Maurílio N. Vieira, *Análise Paramétrica de Sinais de Voz Baseada em Estimação Conjunta do Modelo Fonte-Filtro*. XXX Simpósio Brasileiro de Telecomunicações - SBrT, Brasília, Brazil, 2012.

ISBN: 978-1-61208-249-3