SPEECH ANALYSIS AND SYNTHESIS USING A GLOTTAL EXCITED AR MODEL WITH DTW-BASED GLOTTAL DETERMINATION

G. Cohen and D. Malah

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, Israel. E-mail: gilco@rotem.technion.ac.il

ABSTRACT

In this paper we present a new method for determining the excitation waveform of a glottal excited speech synthesizer. In the glottal excited speech model, voiced speech is decomposed into a parametric glottal shaped excitation signal and an AR vocal tract filter. The new glottal determination method can be used iteratively with vocal tract filter updates in model analysis. In each iteration the new glottal pulse positioning is determined at sub-sample resolution using by a *Multi-Dimensional Dynamic Time Warping (MD-DTW)* algorithm, using an analysis-by-synthesis approach to minimize the squared-error between the original and the synthetic speech.

1. SPEECH SYNTHESIS USING GLOTTAL EXCITATION

Speech synthesis algorithms are currently in wide use and are expected to be much more in demand in the future, since man-machine communication will be based more heavily on speech. Automatic translation, speech coding, voice conversion and modification, text-to-speech and more, all have to use a good speech synthesizer. The speech quality in such applications rely heavily on the choice of the speech synthesis algorithm applied.

A common approach for speech synthesis is based on a *speech production model*. Speech synthesis in this case is equivalent to specifying successive model parameter vectors describing the evolution of speech in time.

In the *Linear Prediction Vocoder*, speech is synthesized using an impulse train or white noise excitation. The synthesized speech quality is often judged as unnatural due to incorrect voicing decisions, poor spectral resolution, and oversimplified excitation functions. Some improved models for the speech production are motivated by the research on the physiology of the human voice production mechanism, and use a glottal excitation waveform. In recent years research has shown that the excitation waveform is individualistic and contains attributes that are associated with vocal quality, vocal disorders, and individual speaking characteristics.

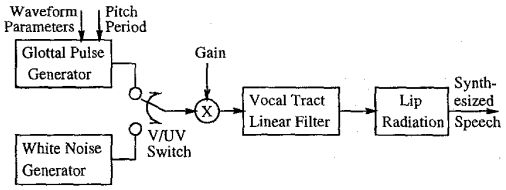


Figure 1: Glottal Excited Speech Synthesizer.

The glottal excited AR synthesizer is based on three elements: A glottal pulse generator; a linear production vocal tract filter; and a lip radiation filter (Fig. 1). The lip radiation effect is modeled as a first order differentiator filter $(1-\mu z^{-1})$, with μ a constant close to one. The vocal tract is modeled usually as an All-Pole linear filter, but can be generalized to any ARMA filter. The glottal volume velocity has an *openning phase* with increasing air flow, a *closing phase* with a sharp decrease in air flow, and an (almost) *closed-phase* with very small (decaying) air flow. Glottal volume velocity shape and timing all vary with time. The glottal pulse generator in the voice production model (Fig. 1) has to capture some or all of these features to produce high quality speech. To simplify the analysis algorithms of such speech models, the order of the vocal tract filter and the differentiator are often reversed. In this way, one can define a glottal excitation waveform model describing the differentiated-volume-velocity.

2. GLOTTAL WAVEFORM ANALYSIS

All glottal waveform analysis methods must assume high quality, linear phase, clear speech recording (especially at low frequencies) for the speech production

model to hold with a high degree of accuracy. Glottal waveform determination methods can be classified into two categories:

- Inverse filtering methods.
- Methods that search the parameter space of a modeled glottal excitation and filter coefficients.

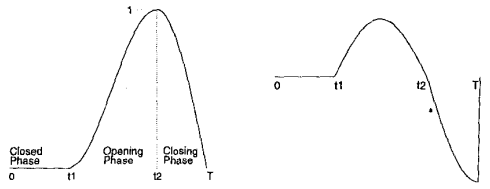
The first class of glottal waveform determination methods is known as inverse filtering. If the transfer function of the vocal tract is known, glottal volume velocity can be determined through deconvolution of the speech waveform. Although many varieties of inverse filtering techniques exist, the vocal tract transfer function is generally estimated over the closed-phase interval of the pitch period to obtain an all-pole model. Classic inverse filtering work was done by Wong, Markel and Gray [1], And some varients on it where done by Alku [2] and recently by Childers [3].

The estimated excitation waveform resulting from the inverse filtering, is then parameterized using a glottal model. Many models for the glottal pulse waveform are known in the literature [4, 5, 6], and they all describe one pitch period of the volume velocity or differentiated volume velocity. One of the most widely used model in inverse filtering is the *Liljencrants-Fant (LF)* model [5] which is expressed in terms of five parameters (including pitch duration).

The second class of analysis algorithms makes a prior assumption of a glottal model. A search is performed in the parameter space of the model and the filter coefficients, for an optimal set which minimizes a pre-defined (weighted) synthesis error. A straight forward method to analyze each pitch period in this case would be to estimate the vocal tract filter coefficients, gain and glottal parameters, combined, using a gradient decent algorithm. In this method, it is difficult to supervise the parameters, assuring for instance, non-negative, time increasing glottal parameters and the stability of the AR filter. Also, the error gradient can not be supplied analytically. Therefore, an update in glottal parameters and vocal tract filter coefficients is done *separately in an iterated manner.*

Because the glottal parameters search is incorporated in the analysis algorithm, the use of more "modest" models (in terms of the number of parameters) is needed for computation efficiency. This does not necessarily means a compromise in synthesized speech quality, because, as opposed to most inverse filtering techniques, interaction between glottal parameters and vocal tract filter coefficients exists while performing the analysis.

Work in this last class of analysis algorithms has been done by Hedelin[7], Fujisaki and Ljungqvist[8] and others. In these and other algorithms, if the squared synthesis error is sought, it leads to non-linear minimization problems in the filter coefficients and glottal parameters. As far as we know, a *weighted squared-error* has always been used to make the error linear in some of the parameters or coefficients. Usually, the weighting function is the denominator of the vocal tract filter, resulting in a weighted error which is linear in the filter coefficients. These weighting functions reduce computation complexity, but rarely agree with the human noise masking characteristic.



$$g(t) = \begin{cases} 0 & ; \quad 0 \le t < t_1 \\ \sin^2(\frac{\pi}{2}\frac{t-t_1}{t_2-t_1}) & ; \quad t_1 \le t < t_2 \\ \cos(\frac{\pi}{2}\frac{t-t_2}{T-t_2}) & ; \quad t_2 \le t < T \end{cases}$$

$$\frac{dg(t)}{dt} = \begin{cases} 0 & ; \quad 0 \le t < t_1 \\ \frac{\pi}{2(t_2-t_1)}\sin(\pi\frac{t-t_1}{t_2-t_1}) & ; \quad t_1 \le t < t_2 \\ -\frac{\pi}{2(T-t_2)}\sin(\frac{\pi}{2}\frac{t-t_2}{T-t_2}) & ; \quad t_2 \le t < T \end{cases}$$

Figure 2: Three parameter glottal model. The volume-velocity model (top left). The differentiated volume-velocity model (top right) used in the iterative analysis algorithm. Mathematical formulations at center and bottom.

3. MULTI-DIMENSIONAL DTW DETERMINATION OF THE GLOTTAL EXCITATION

Our method for determining the glottal excitation can be incorporated in the analysis methods of the second class. We limit ourselves to a family of glottal models, such as the one defined in Fig. 2 [4, 7], which have three phases and are described uniquely by three parameters: The pitch period T, the relative opening duration $(t_2-t_1)/T$ and the relative closing duration $(T-t_2)/T$. All these models assume zero air flow at the closed glottis phase, and as such are simplified versions of more complicated models, such as the LF model.

We will also assume here that we are given an algorithm, which is able to estimate (or update) the vocal tract filter coefficients and gain, for a given glottal excitation (one pitch periond length or more).

Analysis of the speech begins with two processes: First, the speech is segmented into voiced, unvoiced

and silence sections. Unvoiced sections are analyzed in the same manner as in the LPC Vocoder. Second, a coarse marking of glottal closure points in the voiced speech must be made. Glottal closure points are the instances of maximum (negative) excitation values of the differentiated volume velocity, and therefore are very important for the positioning of the modeled glottal waveform. Much research has been done on accurately estimating those points from the speech signal [3]. The pitch period is defined as the duration between successive glottal closure points.

The analysis window length will be the duration of three pitch periods, and we start with one at the beginning of a voiced speech section. Initial glottal excitation waveform parameters are calculated using the closure marks in the window and some averaged opening and closing relative durations [4].

Using those parameters, the continuous time glottal waveform can be defined and sampled at the working sampling rate. To avoid aliasing distortion due to sampling, the sampling rate must be sufficiently high, or else some smoothing of the glottal pulse must be done prior to the sampling. So far we have a three pitch period long initial guess for the glottal excitation in the analysis window. At this point one AR vocal tract filter for the whole analysis window or a different set of filter coefficients for each pitch period in the window can be evaluated. Synthesized speech is the result of the glottal excitation passing through the AR filter, determining an initial SNR value relative to the original speech.

Assuming a known vocal tract filter in the analysis window, we would like to update our initial guess of the glottal excitation in order to increase the SNR. The fact that the glottal waveform is described uniquely by its time points t_1, t_2, T (see Fig. 2), allows us to formulate this problem as a *Dynamic Time Warping (DTW)* problem: mapping those time points from one time-axis to another in order to achieve better synthesis.

We therefore define a DTW plane, with the horizontal time-axis as the current time-axis, and the vertical one as the new desired time-axis. The 45° line in this plane has 10 DTW lattice points on it, three for each of the three glottal waveforms and one in the plane origin. These lattice points represent no change in glottal timing and shape. We allow for small time perturbation steps, $\pm\Delta t_j$, $j = 2, 3, \ldots, 9$ for each of the time points t_j, where j is an index successively labeling these points. The first and last time points ($j = 1, 10$) are not allowed to move, for they define the the analysis window ends. The 45° line and the perturbation points define the DTW lattice, through which an optimal time warping path should be traced. Fig. 3 illustrates such

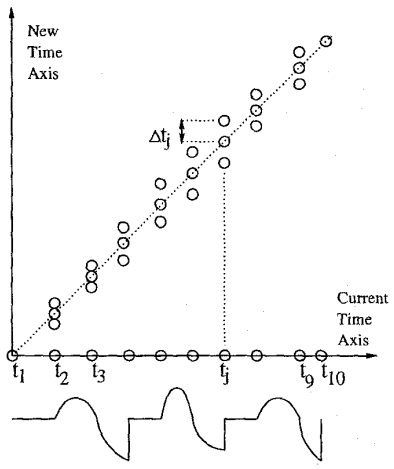


Figure 3: DTW lattice for updating glottal timing.

a possible DTW lattice.

The DTW requires a cost function to be evaluated for each move from one lattice point to another. In order that the best path will define a new glottal waveform that will give minimum squared synthesis error, *the cost measurement should be related to the increase in squared-error*, resulting from passing the (fragment of) glottal waveform associated with these lattice points, through the current vocal tract filter. This requires that each lattice point holds the following:

- The accumulated squared error of the optimal path up to the current lattice point.
- An index describing from which previous lattice point this optimal local path has arrived.
- And finally, the last P samples of the synthesized speech, where P is the order of the AR filter. These are needed as the correct initial state of the filter for further synthesis through this lattice point.

Implementation of the DTW algorithm as described up to now may result in poor performance, i.e., the chosen path through the lattice not being the one to achieve minimum squared synthesis error. This can be explained as follows: In the standard DTW formulation, the cost of moving from one lattice point to the other should depend *only on these points*. Because the vocal tract AR filter has memory, this is not the case here. The amount of increase in squared-error associated with moving form a previous lattice point to the current one, depends not only on the nature of the local

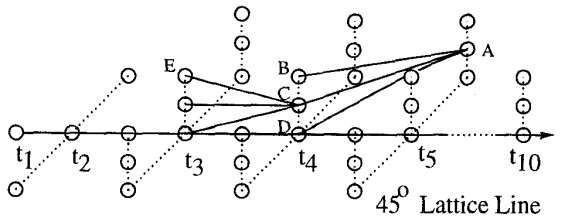


Figure 4: Three levels, 3D-DTW lattice resulting in a delayed decision path.

fragment of glottal excitation, but also on the *filter's initial state.* The initial state is a result of the chosen path up to the previous lattice point.

For an improved solution, we used a *Multi-Dimensional DTW (MD-DTW)* algorithm, as described by Stettiner et al. [9], allowing path control and non-local cost. In this case, a delayed decision path is evaluated through the lattice. We have chosen to use a 3D lattice, as shown in Fig. 4, by adding two lattice levels to the original flat lattice of Fig. 3. The lattice point "level" and stored index, together describe the (locally) best way to arrive at that point from *two time steps backwards.*

For example, let us say lattice point "A" is now under examination. Point "A" is in level two, so possible paths arriving at it would be only from points "B","C" and "D" (the second column in the previous array). Each of these points has a stored filter initial state vector. The proper fragment of glottal excitation is defined, and passed *three times* through the AR filter, with the *three different filter initial states.* The accumulated squared-error of each synthesis result is calculated, and the point "B","C" or "D" which gives the smallest accumulated squared-error is chosen. Let's assume point "C" was chosen, and its stored index indicates arriving form point "E". That means that the path "E"-"C"-"A" is the (locally) optimal path to point "A". The accumulated squared-error, the last P samples defining the initial state, and an index pointing at "C" are saved in lattice point "A". The process is repeated for all lattice points.

In this process, the number of lattice points is (almost) tripled, but the amount of information each lattice point has to hold does not change. *The projection of the multi-dimensional optimal path on the basic plane defines the time warping function.*

After the glottal excitation waveform was updated, we should again update, or recalculate the filter coefficients and gain (either one set for the whole analysis window or three sets - one for each pitch period). These two updates - glottal and filters - should be repeated iteratively a few times. It is obvious that the filter update mechanism should by itself be aiming to minimize the synthesized speech squared-error, or some other related measure (like Maximum Likelihood, assuming a Gaussian model-error). Only in that way, the two iterative updates will result in a consistent decrease in squared-error (increase in SNR).

There is some freedom in the DTW lattice design. Our policy was as follows: If the perturbation steps in the DTW process are kept constant through all the analysis iterations, the squared-error could converge in a very small number of iterations, to a value far from the possible minimum. In order to achieve high resolution (sub-sample), which is needed for maintaining high quality speech, the perturbation steps are varied. At the beginning, the perturbation step values Δt_j^0, $2 \leq j \leq 9$, are set to some small percentage of the local pitch duration $T(j)$ (in seconds):

$$\Delta t_j^0 = \eta T(j) \; ; \; 2 \leq j \leq 9 \; ; \; \eta \ll 1$$

where $\{T(j)\}$ all take one of three possible values. If at any iteration k ($k \geq 1$), the time point t_j^k was perturbated, i.e., received one of the following values:

$$t_j^{k+1} = t_j^k \pm \Delta t_j^k$$

then no change in perturbation step is made for that point ($\Delta t_j^{k+1} = \Delta t_j^k$). If on the other hand, following the DTW, $t_j^{k+1} = t_j^k$, then the resolution is increased:

$$\Delta t_j^{k+1} = \rho \Delta t_j^k \; ; \; 0 < \rho < 1$$

so that in the next iteration this point can be determined more precisely. typical values we used were $\eta = 0.02$ and $\rho = 0.5$.

The lattice structure represented here is with a three point lattice width. One can define a wider lattice of K point width, which results in a K level 3D lattice. However, this will increase the computation burden.

The iteration process can be stopped according to a number of criterions (one of them or a combination): (1) A given number of iterations has been made, (2) Squared-error convergence, and (3) All perturbation step values are below a certain threshold. When this happens, a window update is made. A feature vector of the first pitch period (containing filter coefficients, gain and glottal pulse timing) is saved, and the analysis window slides one pitch period forward in time. As an initial guess, the opening and closing relative times are calculated from the preceding pitch period and an initial glottal excitation with its time points can be specified. This initial guess is expected to be good, because usually, the glottal pulse shape varies slowly from one pitch period to another. Each "old" time point in the window keeps its own perturbation step value through

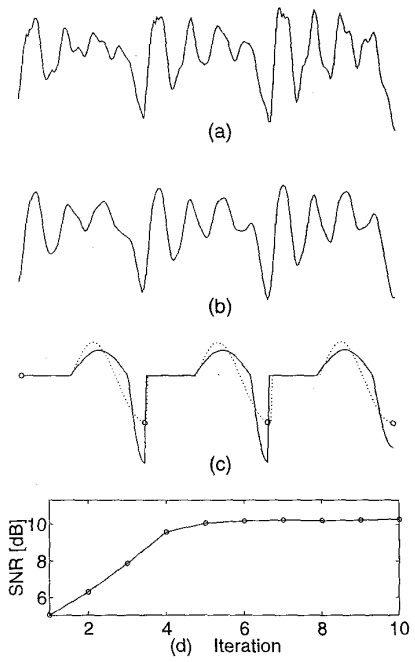


Figure 5: Analysis and Synthesis results. (a) Original Speech. (b) Synthesized Speech using the final glottal excitation. (c) Initial glottal excitation (dotted) with closure points marking. Final glottal excitation (solid). (d) SNR values as a function of DTW iterations.

an analysis window update, and the three new time points are assigned new perturbation step values, as a fixed percent η of the new pitch duration.

4. RESULTS AND DISCUSSION

The MD-DTW method for evaluating the glottal excitation was examined on high quality speech sampled at 8KHz. An example of one analysis frame of voiced male speech is plotted in Fig. 5a. Our initial glottal excitation waveform for this analysis window is plotted in *dotted line* in Fig. 5c. The four initial closure time points are marked on that line. Assuming this initial excitation, an AR filter of order $P = 10$ was calculated separately for each pitch period, in a manner described in [8] (using an AR denominator-weighted squared-error). In order to examine the contribution of the glottal DTW updates to the synthetic speech SNR, successive DTW evaluations where carried out, *without updating the AR filters*. Fig. 5d shows a plot of the SNR value as a function of the DTW iterations. Fig. 5c shows in solid line the final glottal excitation, and Fig. 5b the synthetic speech resulting from the final excitation.

The SNR value increases in the first 7 iterations, and then fluctuates slightly around a final SNR level in further iterations. We do not see a monotonic convergence, because time resolution of the DTW lattice has increased to a point where all possible paths have close squared error, and therefore the approximated MD-DTW solution can no longer choose the optimal one. We see however, that the SNR fluctuations are very small compared to the increase in the overall SNR. Given a constant vocal tract filter, the glottal excitation has adapted itself as best as possible, as seen in the synthesized signal. An improved synthesis of the speech requires an update in the vocal tract filter as well.

We have found that the MD-DTW lattice structure described in Fig. 4 was sufficient, and therefore no extension of the lattice width was made.

We are currently examining efficient iterative algorithms, such as the EM, for maximizing the Likelihood function, in order to iteratively update also the AR coefficients for improved SNR.

5. REFERENCES

[1] D.Y. Wong, J.D. Markel and A.H. Gray, "Least square glottal inverse filtering from the acoustic speech waveform", *IEEE Tran. Acoust. Speech and Signal Proc.* Vol. ASSP-27, No. 4, Aug. 1979, pp. 350-355.

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

[3] D.G. Childers and H.T. Hu, "Speech synthesis by glottal excitation linear prediction", *J. Acoust. Soc. Am.* 96(4), Oct. 1994, pp. 2026-2036.

[4] A.E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels", *J. Acoust. Soc. Am.* 49(2) part 2, 1971, pp. 583-590.

[5] G. Fant, J. Liljencrants, and Q. Lin, "A four parameter model of glottal flow", STL-QPSR 4, 1985, pp. 1-13 .

[6] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for glottal source waveform", *ICASSP*, Tokyo 1986, pp. 1605-1608.

[7] P. Hedelin, "High quality glottal LPC-Vocoding", *ICASSP*, Tokyo 1986, pp. 465-468.

[8] H. Fujisaki and M. Ljungqvist,"Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform",*ICASSP*, 1987, pp. 637-640.

[9] Y. Stettiner, D. Malah and D. Chazan, "Dynamic Time Warping with path control and non-local cost", *12th ICPR*, Jerusalem, Oct. 1994, pp. 174 - 177.