The KlattGrid speech synthesizer

David Weenink

Institute of Phonetic Sciences, University of Amsterdam, The Netherlands

David.Weenink@uva.nl

Abstract

We present a new speech synthesizer class, named KlattGrid, for the Praat program [3]. This synthesizer is based on the original description of Klatt [1, 2]. New aspects of a KlattGrid in comparison with other Klatt-type synthesizers are that a KlattGrid

- is not frame-based but time-based. You specify parameters as a function of time with any precision you like.
- has no limitations on the number of oral formants, nasal formants, nasal antiformants, tracheal formants or tracheal antiformants that can be defined.
- has separate formants for the frication part.
- allows varying the form of the glottal flow function as a function of time.
- allows for any number of formants and bandwidths to be modified during the open phase of the glottis.
- uses no beforehand quantization of amplitude parameters.
- is fully integrated into the freely available speech analysis program Praat [3].

Index Terms: speech synthesizer, KlattGrid, Praat

1. Introduction

A speech synthesizer is an essential tool for speech research. A very well known and widely used speech synthesizer is the Klatt synthesizer [1, 2]. In Fig. 1 we show a schematic diagram of this synthesizer. Since a KlattGrid is based on the same design this is also the diagram of a KlattGrid. The synthesizer consists of four parts:

1. the *phonation part* generates voicing as well as aspiration. It is represented by the top left dotted box labeled with the number 1 in its top right corner.
2. the *coupling part* models coupling between the phonation part and the vocal tract. In the figure it is indicated by the dotted box labeled with the number 2.
3. the *vocal tract part* filters the sound generated by the phonation part. The top right dotted box labeled 3 shows this part as a cascade of formant and antiformant filters. The vocal tract part can also be modeled with formant filters in parallel instead of in cascade.
4. the *frication part* generates frication noise and is represented by the dotted box labeled 4.

A number of implementations of the Klatt synthesizer exist nowadays. However, they all show some, or all, of the limitations of the original design that originates from times that computer memory and processing power were relatively scarce.

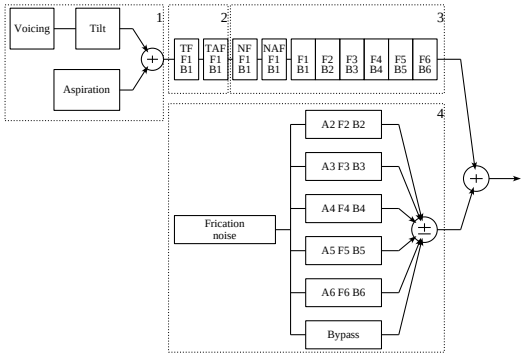

Figure 1: *Schematic diagram of a KlattGrid / Klatt synthesizer with the vocal tract part realized as filters in cascade.*

Necessarily, compromises had to be made at that time in order to achieve reasonable performance.

We present a new speech synthesizer class, a KlattGrid, which is based on the original description of Klatt [1, 2]. There are several new aspects in the KlattGrid in comparison with other Klatt-type synthesizers.

- A Klatt synthesizer is frame-based, i.e. parameters are modeled to be constant during the interval of a frame, typically some five or ten milliseconds. As a consequence, instants of parameter change have to be synchronized on a frame basis. This poses some difficulty in modeling events where timing is important such as a rapidly changing amplitude for plosive bursts. We have removed this limitation by modeling *all* parameters in a KlattGrid as *tiers*. A tier represents a parameter contour as a function of time by (*time*, *value*) points. Parameter values at any time can be calculated from these time stamps by either interpolation or constant extrapolation. For example, a formant frequency tier with two (*time*, *frequency*) points, namely 800 Hz at a time of 0.1 seconds and 300 Hz at 0.3 seconds, is to be interpreted as a formant frequency contour that is constant at 800 Hz for all times before 0.1 seconds, constant at 300 Hz for all times after 0.3 seconds and linearly interpolated for all times between 0.1 and 0.3 seconds (i.e. 675 Hz at 0.15 seconds, 550 Hz at 0.2 seconds, and so on). By leaving the frame-based approach of previous synthesizers, all parameter timings become transparent and only moments of parameter change have to be specified.
- In a Klatt synthesizer one can normally define some six to eight oral formants and one nasal and one tracheal

formant/antiformant pair. In a KlattGrid any number of oral formants, nasal formants and nasal antiformants, tracheal formants and tracheal antiformants are possible.

- In a Klatt synthesizer there is only one set of formant frequencies that has to be shared between the vocal tract part and the frication part. In a KlattGrid the formant frequencies in the frication part and the vocal tract part have been completely decoupled from one another.
- In the Klatt synthesizer the glottal flow function has to be specified beforehand. A KlattGrid allows varying the form of the glottal flow function as a function of time.
- In the Klatt synthesizer only the frequency and bandwidth of the first formant can be modified during the open phase. In a KlattGrid there is no limit to the number of formants and bandwidths that can be modified during the open phase of the glottis.
- In Klatt's synthesizer all amplitude parameters have been quantized to 1 dB levels beforehand. In a KlattGrid there is no such quantization. All amplitudes are represented according to the exact specifications. Quantization only takes place on the final samples of a sound when it has to be played or saved to disk (playing with 16-bit precision, for example). Of course sampling frequencies can be chosen freely.
- A KlattGrid is fully integrated into the speech analysis program Praat [3]. This makes the synthesizer available on the major operating systems of today: Linux, Windows and Mac OS X. At the same time all visualizations and analysis methods of the Praat program become directly available for the synthesized sounds.

More details on the KlattGrid can be found in the following sections which will describe the four parts of the synthesizer in more detail.

2. The phonation part

The phonation part serves two functions:

1. It generates voicing. Part of this voicing are timings for the glottal cycle. The part responsible for these timings is shown by the box labeled "Voicing" in Fig. 1. The start and end times of the open phase of the glottis serve to:
   - generate glottal flow during the open phase of the glottis.
   - generate breathiness, i.e. noise that occurs only during the open phase of the glottis.
   - calculate when formant frequencies and bandwidths change during the open phase (if formant change information is present in the coupling part).
2. It generates aspiration. This part is indicated by the box labeled "Aspiration" in Fig. 1. In contrast with breathiness, aspiration may take place independently of any glottal timing.

The tiers that modify phonation are:

Pitch tier. For voiced sounds the pitch tier models the fundamental frequency as a function of time. Pitch equals the number of glottal opening/closing cycles per unit of time. In the absence of flutter and double pulsing, the pitch tier is the only determiner for the instants of glottal closure.

Voicing amplitude tier. The voicing amplitude regulates the maximum amplitude of the glottal flow in dB SPL. The reference amplitude at 0 dB is 20 μPa. This means that a flow with amplitude 1 corresponds to $20 \log(1/(20 \cdot 10^{-6})) \approx 94$ dB. To produce a voiced sound it is essential that this tier is not empty.

Flutter tier. Flutter models a kind of "random" variation of the pitch and is input as a number from zero to one. This random variation can be introduced to avoid the mechanical monotonic sound whenever the pitch remains constant during a longer time interval. The fundamental frequency is modified by a flutter component according to the following semi-periodic function that we adapted from [2]: $F_0'(t) = 0.01 \cdot \text{flutter} \cdot F_0(\sin(2\pi 12.7t) + \sin(2\pi 7.1t) + \sin(2\pi 4.7t))$

Open phase tier. The open phase tier models the open phase of the glottis with a number between zero and one. The open phase is the fraction of one glottal period that the glottis is open. The open phase tier is an optional tier, i.e. if no points are defined then a sensible default for the open phase is taken (0.7).

Power1 and power2 tiers. These tiers model the form of the glottal flow function during the open phase of the glottis as $\text{flow}(t) = t^{\text{power1}} - t^{\text{power2}}$, where $0 \leq t \leq 1$ is the relative time that runs from the start to the end of the open phase. If these tiers have no values defined default values power1=3 and power2=4 are used.

Collision phase tier. The collision phase parameter models the last part of the flow function with an exponential decay function instead of a polynomial one. A value of 0.04, for example, means that the amplitude will decay by a factor of $e \approx 2.7183$ every 4 percent of a period.

Spectral tilt tier. Spectral tilt represents the extra number of dB's the voicing spectrum should be tilted down at 3000 hertz [1]. This parameter is necessary to model "corner rounding", i.e. when glottal closure is non simultaneous along the length of the vocal folds. If no points are defined in this tier, spectral tilt defaults to 0 dB and no spectral modifications are made.

Aspiration amplitude tier. The aspiration amplitude tier models the (maximum) amplitude of noise generated at the glottis. The aspiration noise amplitude is, like the voicing amplitudes, specified in dB SPL. This noise is independent of glottal timings and is generated from random uniform noise which is filtered by a very soft low-pass filter.

Breathiness amplitude tier. The breathiness amplitude tier models the maximum noise amplitude during the open phase of the glottis. The amplitude of the breathiness noise is modulated by the glottal flow. It is specified in dB SPL.

Double pulsing tier. The double pulsing tier models diplophonia (by a number from zero to one). Whenever this parameter is greater than zero, alternate pulses are modified. A pulse is modified with this single parameter tier in two ways: it is delayed in time and its amplitude is attenuated. If the double pulsing value is maximum ($= 1$), the time of closure of the first peak coincides with the opening time of the second one (but its amplitude will be zero).

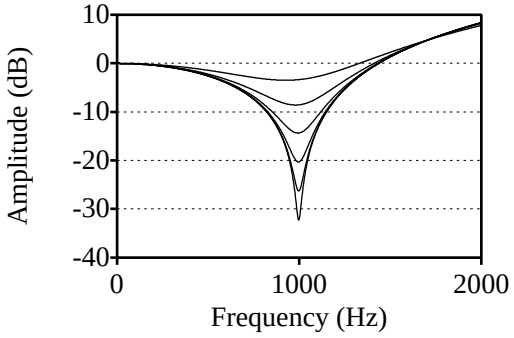

Figure 2: *Example of frequency responses of formant/antiformant pairs.*

3. The vocal tract part

The sound generated by the phonation part of a KlattGrid may be modified by the filters of the vocal tract part. These filters are the oral formant filters, nasal formant filters and nasal antiformant filters. A formant filter boosts frequencies and an antiformant filter attenuates frequencies in a certain frequency region. For speech synthesis these filters can be used in cascade or in parallel. Each formant filter is governed by two tiers: a formant frequency tier and a formant bandwidth tier. In case of parallel synthesis an additional formant amplitude tier must be specified. Formant filters are implemented in the standard way as second order recursive digital filters of the form $y_n = ax_n + by_{n-1} + cy_{n-2}$ as described in [2] (x_i represents input and y_j output). These filters are also called digital resonators. The coefficients b and c at any time instant n can be calculated from the formant frequency and bandwidth values of the corresponding tiers. The a parameter is only a scaling factor and is chosen as $a = 1 - b - c$; this makes the frequency response equal to 1 at zero frequency. Antiformants are second order filters of the form $y_n = a'x_n + b'x_{n-1} + c'x_{n-2}$. The coefficients are also determined as described in [2].

As an example we show in fig. 2 the frequency responses of formant/antiformant pairs where both formant and antiformant have the same "formant" frequency, namely 1000 Hz, but different bandwidths. The bandwidth of the antiformant filter was fixed at 25 Hz but the bandwidth of the formant filter doubles at each step. From top to bottom it starts at 50 Hz and then moves to 100, 200, 400 and 800 Hz values. A perfect spectral "dip" results without hardly any side-effect on the spectral amplitude. Best spectral dips are obtained when the formant bandwidth is approximately 500 Hz. For larger bandwidths the dip will not become any deeper, the flatness of the spectrum will disappear and especially the higher frequencies will be amplified substantially.

4. The coupling part

The coupling part of a KlattGrid models the interactions between the phonation part, i.e. the glottis, and the vocal tract. Coupling is only partly shown in Fig. 1, only the tracheal formants and antiformants are shown. We have displayed them in front of the vocal tract part after the phonation part because tracheal formants and antiformants are implemented as if they filter the phonation source signal.

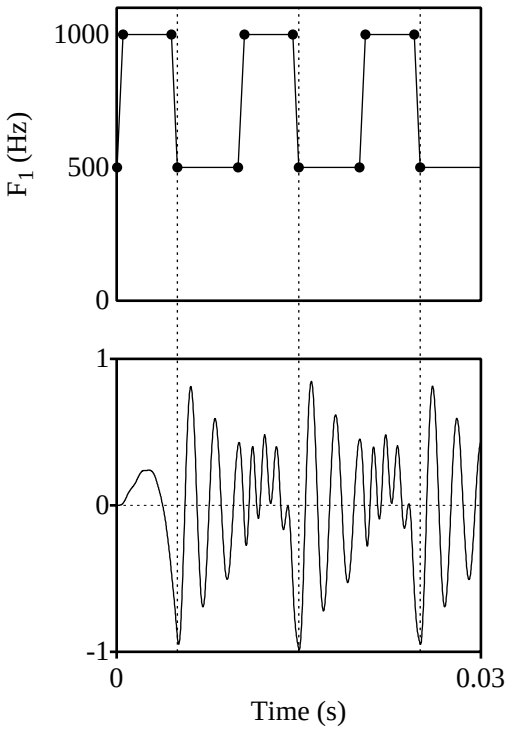

Figure 3: *Synthesized example of extreme coupling between the vocal tract and the glottis. The frequency of the first oral formant is 500 Hz and increases by 500 Hz during each open phase of the glottis (pitch is 100 Hz, open phase is 0.5, and bandwidth constant at 50 Hz).*

Besides the tracheal system with its formants and antiformants the coupling part also models the change of formant frequencies and bandwidths during the open phase of the glottis. With a so-called *delta formant grid* we can specify the amount of change of any formant and/or bandwidth during the open phase of the glottis. The values in the delta tiers will be added to the values of the corresponding formant tiers but *only during the open phase of the glottis*.

In Fig. 3 we show an example where an extreme coupling value has been used for a clear *visual* effect. We have generated a voiced sound with a 100 Hz pitch, an open phase of 0.5 to make the duration of the open and closed phase equal, and only one formant. In the figure the formant bandwidth is held constant at 50 Hz while its frequency is modified during the open phase. The oral formant frequency is set to 500 Hz. By setting a delta formant point to a value of 500 Hz we accomplish that during the start of the open phase of the glottis, the formant frequency will increase by 500 Hz to 1000 Hz. At the end of the open phase it will then decrease to the original 500 Hz value of the formant tier. To avoid instantaneous changes we let the formant frequency increase and decrease with the delta value in a short interval that is one tenth of the duration of the open phase. The top display shows the first formant frequency as a function of time during the first 0.03 s. This is exactly the duration of three pitch periods; the moments of glottal closure are indicated by dotted lines. The bottom display shows the corresponding one-formant sound signal. The 100 Hz periodicity is

visible as well as the formant frequency doubling in the second part of each period: we count almost two and a half periods of this formant in the first half of a period, the closed phase, and approximately five during the second half of a period, the open phase.

5. The frication part

The frication part is an independent section in the synthesizer which gives the opportunity to add the frication noise completely independent of the phonation and the vocal tract part. The frication sound is added to the output of the vocal tract part. A layout of the frication part is shown at the bottom of Fig. 1 in the dotted box labeled 4. The following tiers specify the frication sound:

Frication amplitude tier. This tier regulates the maximum amplitude of the noise source in dB SPL before any filtering takes place. In Fig. 1 this part is represented by the rectangle labeled "Frication noise". This noise source is uniformly distributed random noise.

Formant frequency and bandwidth tiers. To shape the noise spectrum a number of parallel formant filters are available whose frequencies, bandwidths and amplitudes can be specified. In the figure we have limited the number of formants to five but in principle this number is not limited.

Formant amplitude tiers. Each formant is governed by a separate amplitude tier with values in dB. These formant amplitudes act like multipliers and may amplify or attenuate the formant filter input. For formant amplitudes 0 dB means an amplification of 1. Formants can be increased by giving positive dB values and decreased by giving negative values.

Bypass tier. The bypass tier regulates the amplitude of the noise that bypasses the formant filters. This noise is added straight from the noise source to the output of the formant filters. The amplitude is in dB's, where 0 dB means a multiplier of 1.

6. A KlattGrid scripting example

At the bottom of this section we show a simple script to synthesize a diphthong. This script can be run in Praat's script editor. The first line of the script creates a new KlattGrid, named "kg", with start and end times of 0 and 0.3 s, respectively. The rest of the parameters on this line specify the number of filters to be used in the vocal tract part, the coupling part and the frication part and are especially important for now (additional filters can always be added to a KlattGrid).

The second line defines a pitch point of 120 Hz at time 0.1 s. The next line defines a voicing amplitude of 90 dB at time 0.1 s. Because we keep voicing and pitch constant in this example the exact times for these points are not important, as long as they are in the domain on which the kg KlattGrid is defined. With the pitch and voicing amplitude defined, there is enough information in the KlattGrid to produce a sound and we can now Play the KlattGrid (line 4). During 300 ms you will hear the sound as produced by the glottal source alone. This sound normally would be filtered by a vocal tract filter. But we have not defined the vocal tract filter yet (in this case the vocal tract part will not modify the phonation sound).

In lines 5 and 6 we add a first oral formant with a frequency of 800 Hz at time 0.1 s, and a bandwidth of 50 Hz also at time 0.1 s. The next two lines add a second oral formant at 1200 Hz with a bandwidth of 50 Hz. If you now play the KlattGrid (line 9), it will sound like the vowel /a/, with a constant pitch of 120 Hz. Lines 10 and 11 add some dynamics to this sound; the first and second formant frequency are set to the values 300 and 600 Hz of the vowel /u/; the bandwidths have not changed and stay constant with values that were defined in lines 6 and 8. In the interval between times 0.1 and 0.3 s, formant frequency values will be interpolated. The result will now sound approximately as an /au/ diphthong.

This script shows that with only a few commands we already may create interesting sounds.

```
1  Create KlattGrid... kg 0 0.3 6 1 1 6 1 1 1
2  Add pitch point... 0.1 120
3  Add voicing amplitude point... 0.1 90
4  Play
5  Add oral formant frequency point... 1 0.1 800
6  Add oral formant bandwidth point... 1 0.1 50
7  Add oral formant frequency point... 2 0.1 1200
8  Add oral formant bandwidth point... 2 0.1 50
9  Play
10 Add oral formant frequency point... 1 0.3 300
11 Add oral formant frequency point... 2 0.3 600
12 Play
```

7. Conclusions

We have described the KlattGrid class, a new variant of the Klatt synthesizer. In a KlattGrid many of the limitations of the original design have been removed. The most important new feature is that a KlattGrid is time-based instead of frame-based. This makes it easier to specify events where timing is critical. A KlattGrid has no limitations on the possible number of oral, nasal or tracheal formants and antiformants. Instead of being fixed, the form of the glottal flow function can be modified as well. Amplitude values are not quantized beforehand and sampling frequencies can be chosen freely. Being incorporated into the speech analysis and synthesis program Praat makes it freely available for the major computer platforms of today and has the additional benefit of the complete visualization, scripting and analysis environment of Praat.

8. Acknowledgements

The KlattGrid project was carried out within the STEVIN programme which was funded by the Dutch and Flemish Governments (http://www.stevin-tst.org).

9. References

[1] Klatt, D.H. and Klatt, L.C., "Analysis, synthesis, and perception of voice quality variations among female and male talkers", J. Acoust. Soc. Am., 87:820–857, 1990.

[2] Klatt, D.H.,"Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Am. 67:971–995, 1980.

[3] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer (Version 5.1.07), [Computer program]", http://www.praat.org/, 2009.