# Formant-based synthesis of singing

*Sten Ternström, Johan Sundberg*

Department of Speech, Music and Hearing, School of Computer Science and Communication, Kungliga Tekniska Högskolan, SE-100 44, Sweden

www.speech.kth.se

## Abstract

Rule-driven formant synthesis is a legacy technique that still has certain advantages over currently prevailing methods. The memory footprint is small and the flexibility is high. Using a modular, interactive synthesis engine, it is easy to test the perceptual effect of different source waveform and formant filter configurations. The rule system allows the investigation of how different styles and singer voices are represented in the low-level acoustic features, without changing the score. It remains difficult to achieve natural-sounding consonants and to integrate the higher abstraction levels of musical expression.

**Index Terms**: formant synthesis, singing

## 1. Background

Singing synthesis at KTH has its roots in the 1970's, when Sundberg and Gauffin modified the text-to-speech systems developed by Carlson and Granström. An analogue singing synthesiser called MUSSE was built by Larsson in 1977 [1]. It included vibrato and other song-specific features, and could be played with a piano keyboard and joystick, or be remote-controlled by a minicomputer running a rule system. In the 1990's, several digital implementations of MUSSE were made by Ternström and Berndtsson [2]. The synthesis model described here is a descendant of these, built with Aladdin, a commercial DSP tool that was another outcome of this work (Aladdin Interactive DSP 3.0, Hitech Development AB, Täby, Sweden).

## 2. Rule system

The rule system is based on a legacy version of Rolf Carlson's RULSYS, a FORTRAN program in DOS, but we hope to modernise this any decade now, by merging the pronunciation rules into our more recent Director Musices software. The original text-to-speech rules (for Swedish) were extensively modified for singing, and rules for the musical performance have been added. RULSYS compiles the script of rules, and also a singer definition script, and then renders the score (with lyrics and melody) into a parameter file. There are currently 28 integer parameters, which are updated 100 times per second. The parameter file is transferred to the DSP-resident synthesis model from the host PC.

## 3. Synthesis model

A block diagram of the synthesis model is shown in Figure 1.The signal that controls the nominal F0 is perturbed by white noise, and then smoothed with a moderately resonant filter at about 4 Hz. This simulates both irregular flutter and smoothed F0 transitions with adjustable overshoot [3]. To this an adjustable sinusoidal vibrato is added. The vibrato cycle is not aligned with the note boundaries.

The widely studied LF model [4] of the source waveform is not used, because it is awkward to implement under the multiple constraints of high precision in F0, minimal aliasing and of the Aladdin run-time library[1]. Instead, the source oscillator produces a train of sinc pulses, $\sin(x)/x$, with a flat spectrum. The pulse excitations are independent of the sampling interval. Each pulse is windowed to the glottal period time T0 with a Hanning-like window. Even for small values of the glottal period time T0 (high values of F0), aliasing with this window is insignificant, at 16 kHz sampling rate. The reasons for using such a low sampling rate are partly the desire to synthesise in real time on legacy hardware, and partly that raising the sampling rate would necessitate a departure from our standard tract configuration with eight formants.

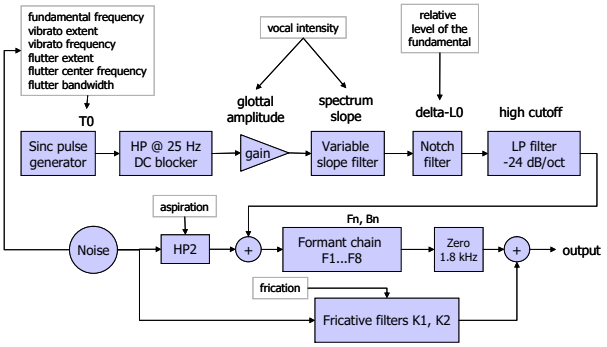


Figure 1. *Block diagram of the current KTH formant synthesis model.*

To approximate the glottal pressure waveform, the sinc oscillator is followed by four filters: (1) a DC blocker, being a first order high-pass filter at 25 Hz; (2) a variable slope filter, with a cutoff fixed at 100 Hz and a slope adjustable from –12 to 0 dB per octave in 0.01 dB increments; (3) a notch filter whose resonance frequency follows F0 so as to give control over the relative level ΔL0 of the fundamental partial only, from –20 to +20 dB; and finally (4) a fourth-order variable low-pass Butterworth filter that is used to attenuate further the high end of the source spectrum. An example source spectrum is shown in Figure 2.

The finished source signal is fed into a chain of formant filters, F1…F8. The model has no nasal branch. An almost gaussian noise generator feeds a fricative branch with two resonance filters. The same noise is used for aspiration and for randomisation of F0 flutter.

Conventionally, formant synthesis more or less stops at 4-5 kHz. Here, a recent improvement to the synthesis is the grouping of formants F6-F8 at 5800, 6500 and 7100 Hz, with bandwidths of 300, 300 and 370 Hz. This creates a cluster around 6500 Hz which mimics a similar cluster that is often found in loud singing voice, but some 40-50 dB below the

main spectrum peak (Figure 3). To make the spectral level of this cluster a better match for live data, an inverse low-pass filter is added at about 1800 Hz. The treble end of the spectrum then improves markedly, especially for male voices, although care must be taken to avoid a buzzy timbre. This high-frequency cluster is of course an approximation. At frequencies above 5 kHz or so, the notion of distinct formant frequencies is no longer appropriate, because of the increasing profusion of higher-order resonance modes in the vocal tract. On the other hand, the ear's critical bands in this region are 1 kHz wide or more, so detail is probably less important.

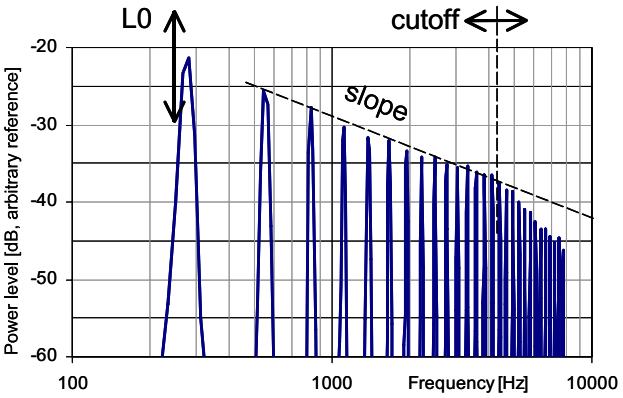


Figure 2. *The source spectrum and the options for controlling its shape*

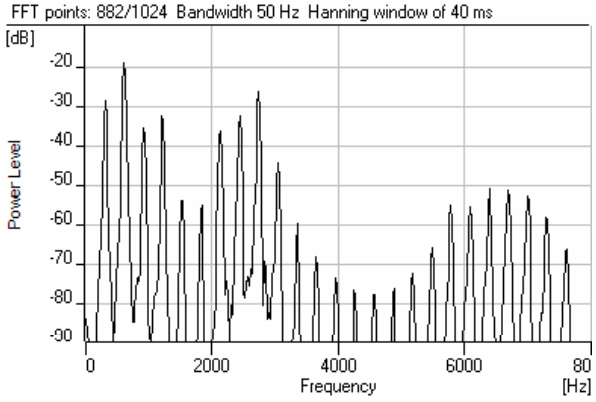


Figure 3. *Spectrum of a synthesized baritone* [a] *vowel, with a weak but readily perceived peak around 6500 Hz.*

## 4. Singer and style characteristics

The main advantage of source-filter synthesis over concatenating synthesis is that it stimulates the search for performance principles and regularities. In addition, it allows modification of voice characteristics without resorting to a huge database of sounds. On the other hand, it is difficult to make consonants and transitions sound very natural. We give four examples, all synthesised from the same excerpt of a score, but with minor modifications to the performance rules and to the singer properties.

1) Operatic baritone. This is the default rule set. A new control of spectrum slope was implemented for this example.
2) Mixed: Formant settings obtained from a female singer reportedly performing in a mode that is intermediate between opera and musical theatre.
3) Male jazz club singer. More aspirative noise. Vibrato is added only toward the end of long tones.
4) Child singer. The sampling rate was raised to 22050 Hz and a simpler source pulse with a steeper spectrum roll-off was used. The higher formants were scaled up about 1.5 times and more aspirative noise was added. The formant bandwidths were doubled, and the vibrato was removed.

Clearly, more effort could be invested in the rules and singer properties, but the examples still serve to illustrate some of the variation that is possible, with very modest storage requirements. The RULSYS program uses only 500 Kb of disk space including rules, scripts and compiler files. The synthesis model scripts run to about 15 Kb of text. The DSP-resident code for the real-time synthesis runs in less than 36K words of DSP memory, including its run-time library, operating kernel, and memory buffers. The synthesiser can also be controlled and inspected interactively from the host computer or via the MIDI protocol.

## 5. Note

[1] It may be noted in passing that, because the LF pulse is essentially the truncated response of an unstable second-order filter, which can be thought of as an anti-causal stable filter, it is possible to approximate a *time-reversed* LF source, by having sinc pulses excite a suitably tuned second-order filter. This topic, as well as the remaining problem of implementing the return phase, have been treated by [5].

## 6. References


[1] Larsson, B., "Music and Singing Synthesis Equipment (MUSSE)", STL-QPSR, 18(1), 38-40, 1977. http://www.speech.kth.se/prod/publications/files/qpsr/1977/1977_18_1_038-040.pdf

[2] Berndtsson, G. and Sundberg, J. "The MUSSE DIG singing synthesis." In A Friberg, J Iwarsson, E Jansson & J Sundberg (eds.), SMAC 93 (Proceedings of the Stockholm Music Acoustics Conference 1993), Stockholm, Roy. Sw. Academy of Music, Publ. No 79, 279-281, 1994.

[3] Ternström, S., and Friberg, A. "Analysis and simulation of small variations in the fundamental frequency of sustained vowels", STL-QPSR 30 (3), 1-14, 1989. http://www.speech.kth.se/prod/publications/files/qpsr/1989/1989_30_3_001-014.pdf

[4] Fant, G., Liljencrants, J., and Lin, Q.G., "A four-parameter model of glottal flow", STL-QPSR 26 (4):1-13, 1985. Available online at http://www.speech.kth.se/prod/publications/files/qpsr/1985/1985_26_4_001-013.pdf

[5] Doval, B., d'Alessandro, C., and Henrich, N. "The voice source as a causal/anticausal linear filter", Workshop proceedings, VOQUAL'03, Geneva, August 27-29, 2003.