Concatenation-based MIDI-to-Singing Voice Synthesis

Michael W. Macon
Dept. of Electrical Engineering
Oregon Grad. Inst. of Science and Tech.
P.O. Box 91000, Portland, OR 97291-1000
macon@ee.ogi.edu

Leslie Jensen-Link
Momentum Data Systems
6195 Heards Creek Dr., N.W.
Atlanta, GA 30328
leslie@mds.com

James Oliverio
Department of Music
Georgia Institute of Technology
Atlanta, GA 30332
james.oliverio@arch.gatech.edu

Mark A. Clements
School of ECE
Georgia Institute of Technology
Atlanta, GA 30332-0250
clements@ee.gatech.edu

E. Bryan George
DSPS R&D Center
Texas Instruments Incorporated
P.O. Box 655303, M/S 8374
Dallas, TX 75265
ebg@ti.com

Abstract

In this paper, we propose a system for synthesizing the human singing voice and the musical subtleties that accompany it. The system, LYRICOS, employs a concatenation-based text-to-speech method to synthesize arbitrary lyrics in a given language. Using information contained in a regular MIDI file, the system chooses units, represented as sinusoidal waveform model parameters, from an inventory of data collected from a professional singer, and concatenates these to form arbitrary lyrical phrases. Standard MIDI messages control parameters for the addition of vibrato, spectral tilt, and dynamic musical expression, resulting in a very natural-sounding singing voice.

1 Background

Singing voice synthesis

Many music synthesis methods have been applied to analysis and resynthesis of selected voice sounds like sustained vowels or consonant-vowel syllables, including formant synthesis [1], wavetables [2], and FM synthesis [3, 4]. Controllable articulatory models such as the SPASM system [5, 6, 7] have also been applied in this area. In recent work [8, 9] the use of *concatenation-based* synthesis of voice has begun to be explored.

The system described in this paper, called LYRICOS, allows for arbitrary lyric input and a fine degree of dynamic control of the voice. In addition, it is able to synthesize a voice that captures voice identity characteristics of the particular human subject used to derive synthesis parameters.

Concatenation-based TTS

One commonly used technique for synthesis of the speech waveform in text-to-speech synthesis is concatenation of short speech units taken from a prerecorded inventory. After concatenation, these units are modified in duration and "melody" to smoothly join each other and achieve the prosody of a natural utterance. In order to perform these modifications without introducing unnatural-sounding artifacts, signal modeling techniques, such as the popular PSOLA technique [10], must be employed. Sinusoidal signal models have been shown to be useful for speech modification [11, 12], speech synthesis [13, 14], and music synthesis [15], among other applications.

ABS/OLA sinusoidal model

The *Analysis-by-Synthesis, Overlap-Add* sinusoidal model [15] provides an attractive framework for speech and music synthesis due to its efficient overlap-add synthesis algorithm and its high quality signal modification capabilities. In the ABS/OLA model, the input signal $s[n]$ is represented by a sum of overlapping short-time signal frames $s_k[n]$.

$$s[n] = \sigma[n] \sum_k w[n - kN_s] s_k[n] \tag{1}$$

where N_s is the frame length, $w[n]$ is a window function that is nonzero over the interval $[-N_s, N_s]$, $\sigma[n]$ is a slowly time-varying gain envelope, and $s_k[n]$ represents the kth frame "synthetic contribution" to the synthesized signal. Each signal contribution $s_k[n]$ is represented as the sum of a small number of constant-frequency sinusoidal components, given by

$$s_k[n] = \sum_{l=0}^{L-1} A_l^k \cos(\omega_l^k n + \phi_l^k) \tag{2}$$

where L is the number of sinusoidal components in the frame, and A_l^k,ω_l^k, and ϕ_l^k are the kth frame sinusoidal amplitudes, frequencies, and phases, respectively. An iterative analysis-by-synthesis procedure is performed to find the optimal parameters for each signal frame, based on a mean-squared error criterion [15].

Overlap-add synthesis is performed by a procedure that uses the inverse fast Fourier transform to compute each contribution $s_k[n]$, rather than sets of oscillator functions, as in other sinusoidal models [12]. Time-scale modification is achieved by changing the time evolution rate of the model parameters for each frame $s_k[n]$ and changing the frame duration, while imposing a "quasi-harmonic" structure on the sinusoidal components to maintain general waveform shape characteristics. Pitch modification is performed within this same context by altering the component frequencies, phases, and amplitudes in such a way that the fundamental frequency is modified while (in the speech case) the formant structure is maintained.

2 LYRICOS System Overview

The system developed in this work, called LYRICOS, is shown in block diagram form in Figure 1. It uses a commercially-available, MIDI-based composition software as an interface for users to specify a musical score, lyrics, and other musically-interesting control parameters such as vibrato and vocal intensity. This control information is stored in a standard MIDI file format that contains all information necessary to synthesize the vocal passage.

Based on this input MIDI file, the system selects synthesis model parameters from an inventory of sinusoidal model voice data. Units are selected to represent segmental phonetic characteristics of the utterance, including coarticulation effects caused by the context of each phoneme. Algorithms described in [14] are then applied to the modeled segments to remove discontinuities in the signal at the joined boundaries. The sinusoidal model is then used to modify the pitch, duration, and spectral characteristics of the concatenated voice units as specified by the input musical score and MIDI control information and synthesize the output.

3 Lyric Synthesis

Voice data collection

To create the voice data corpus for the system, a classically-trained male vocalist [1] was asked to sing 500 nonsense words designed to cover a broad class of coarticulation effects on each vowel [16]. He was instructed to sing half of these words at an arbitrary "high" pitch and the other half at a "low" pitch, and to use very little vibrato. The boundaries of each phoneme within the corpus were then labeled manually. After trimming silences, this resulted in about 10 minutes of singing voice data sampled at 16 kHz. The ABS/OLA sinusoidal model analysis procedure was then applied to create the synthesis inventory.

Variable-length unit selection

A decision-tree algorithm was designed to find the sequence of inventory units with the best match to the "target" specified by the MIDI input. Use of long sequences of phonemes that exactly match the input (for example, finding "ello w" when synthesizing lyrics "hello world")

[1] Thanks to Fay Salvaras of RKM Studios (Atlanta, GA) and Matthew Link.

is desirable, since it minimizes the number of units that must be concatenated. The algorithm was designed to choose strings of one to several phonemes that matched the input phoneme labels, but while simultaneously finding units with a phonetic context, pitch, and duration that matched the MIDI targets. This was accomplished by assigning a "cost" to each unit, with lower costs for short units or poor context match, and a weighted cost proportional to pitch and duration mismatch.

In the LYRICOS algorithm, units were selected independently for each vowel in the musical phrase, moving left to right. Consonants were selected to "fill in the blanks" during a second pass. Approaches to unit selection in speech synthesis, which also use a much larger inventory, generally employ a dynamic programming search to choose the best sequence of units. In future work, we plan to explore such an approach to improve upon the current scheme.

Concatenation/smoothing

Once a sequence of units has been chosen, it is necessary to concatenate them and smooth perceptible discontinuties at the segment boundaries. Figure 2 represents the concatenation of modeled segments within the sinusoidal model framework. Algorithms described in [14] are employed to smooth differences in amplitude, frequency, and phase characteristics of the modeled segments, including smoothing of the spectral envelope (i.e., formant) evolution in the neighborhood of the join. The use of a frequency-domain synthesis model enables a wide range of possibilities.

4 MIDI-based Dynamic Control

Rhythmic characteristics

A natural "quantal unit" of rhythm in both speech and vocal music is the *syllable*—each syllable of lyric is associated with one or more notes of the melody. In order to create synthetic vocals that follow the MIDI-specified rhythmic patterns of the music, it is necessary to find an anchor point within each syllable, and align the "beat" with this anchor.

In speech perception literature, the notion of a syllabic "perceptual center" (PC) has been introduced [17]. In perceptual experiments, multiple listeners were asked to align clicks and speech syllables such that the click–speech–click... intervals were isochronous. In these experiments, it was found that listeners reliably placed the PC at the *vocalic onset* of the syllable (e.g., the beginning of the "o" in the syllable "spoke") [18]. This effect has been noted in the study of rhythmic characteristics of singing [19] as well.

LYRICOS employs rules that align the beginning of the first note in a syllable with the onset of the vowel in that syllable. A simple model for scaling durations of syllables is used. First an average time scaling factor for the syllable, ρ_{syll}, is computed as the ratio:

$$\rho_{syll} = \frac{\sum_{n=1}^{N_{notes}} D_n}{\sum_{m=1}^{N_{phon}} D_m}, \tag{3}$$

where the values D_n are the desired durations of the N_{notes} notes associated with the syllable and D_m are the durations of the N_{phon} phonemes extracted from the inventory to compose the desired syllable. If $\rho_{syll} > 1$, then the vowel in the syllable is *looped* by repeating a set of frames extracted from the stationary portion of the vowel, until $\rho_{syll} \approx 1$. This preserves the duration of the consonants, and avoids unnatural time-stretching effects in plosives. If $\rho_{syll} < 1$, the entire syllable is compressed in time by setting the time-scale modification factor ρ for all frames in the syllable equal to ρ_{syll}.

A more sophisticated approach to the problem would involve phoneme- and context-dependent rules for scaling phoneme durations within each syllable to more accurately represent the manner in which humans perform this adjustment (e.g., [18]).

Pitch variation

Since the prosody modification step in the sinusoidal synthesis algorithm transforms the pitch of every frame to match its MIDI-specified target, the result is a signal that does not exhibit the natural pitch fluctuations of the human voice.

In [20], a simple equation for "quasirandom" pitch fluctuations in speech is proposed:

$$\Delta F_0 = \frac{F_0}{100} \left(\sin(12.7\pi t) + \sin(7.1\pi t) + \sin(4.7\pi t)\right)/3. \tag{4}$$

The addition of this fluctuation to the desired pitch contour gives the voice a more "human" feel, since a slight aperiodic wavering is present. Bennett and Rodet [1] propose a similar model. A global scaling of ΔF_0 is incorporated as a parameter controllable by the user, so that more or less fluctuation can be synthesized.

Abrupt transitions of one note to another at a different pitch are also not a natural phenomena. Rather, singers tend to transition somewhat gradually from one note to another. This effect is implemented within Lyricos by applying a smoothing at note-to-note transitions (*portamento*) in the target pitch contour. Timing of the pitch change by human vocalists is usually such that the transition between two notes takes place *before* the onset of the second note, rather than dividing evenly between the two notes [19]. Thus, the vocalic onset of a syllable corresponds to a stable, already-transitioned pitch.

Vibrato

Trained vocalists produce a 5–6 Hz near-sinusoidal vibrato. As mentioned, pure frequency modulation of the glottal source can represent many of the observed effects of vibrato, since amplitude modulation will automatically occur as the partials "sweep by" the formant resonances. This effect was easily implemented within the sinusoidal model framework by adding a sinusoidal modulation to the target pitch of each note. Vocalists usually are not able to vary the *rate* of vibrato, but rather modify the *modulation depth* to create expressive changes in the voice [5]. Using the graphical MIDI-based input to Lyricos, users can draw contours that control vibrato depth over the course of the musical phrase, thus providing a mechanism for adding expressiveness to the vocal passage. A global setting of the vibrato rate is also possible.

Vocal effort dynamics

Simply scaling the overall amplitude of the signal to produce changes in loudness has the same perceptual effect as turning the "volume knob" of an amplifier; it is quite different from a change in *vocal effort* by the vocalist. Nearly all studies of singing voice mention the fact that the downward tilt of the vocal spectrum increases as the voice becomes softer (e.g., [1, 5, 19]). This effect is conveniently implemented in a frequency-domain representation such as the sinusoidal model, since scaling of the sinusoid amplitudes can be performed. In LYRICOS, an amplitude scaling function based on the work in [1] is used:

$$G_{dB} = \frac{T_{in} \log_{10}(F_l/500)}{\log_{10}(3000/500)}, \tag{5}$$

where F_l is the (Hz) frequency of the lth sinusoidal component and T_{in} is a spectral tilt parameter controlled by a MIDI "vocal effort" control function input by the user. This function produces a frequency-dependent gain scaling function parameterized by T_{in}, as shown in Figure 3.

In studies of acoustic correlates of perceived voice qualities [21, 20], it has been shown that utterances perceived as "soft" and "breathy" also exhibit a higher level of high frequency aspiration noise than fully phonated utterances, especially in females.

In other work with the ABS/OLA model, it was shown that a frequency-dependent noise-like character could be introduced into the signal by employing a subframe phase randomization method [22]. In LYRICOS, this capability has been used to model aspiration noise. The degree to which the spectrum is made noise-like is controlled by a mapping from the MIDI-controlled vocal effort parameter to the amount of phase dithering introduced.

Informal experiments with mapping the amount of randomization to (*i*) a cutoff frequency above which phases are dithered, and (*ii*) the scaling of the amount of dithering within a fixed band, have been performed. Employing either of these strategies results in a more breathy, soft voice, although careful adjustment of the model parameters is necessary to avoid an unnaturally noisy quality in the output. A refined model that more closely reflects the phonetics of loudness scaling and breathiness in singing is a topic for more extensive study in the future.

Vocal tract length scaling

In synthesis of low bass voices using a voice inventory recorded from a baritone vocalist, it was found that the voice took on an artificial-sounding buzzy quality. Through analysis of a simple tube model of the human vocal tract, it can be shown that the nominal formant frequencies associated with a longer vocal tract are lower than those associated with a shorter vocal tract [23]. Because of this, larger people usually have voices with a "deeper" quality; bass vocalists are typically males with voices possessing this characteristic.

In LYRICOS, we approximate the differences in vocal tract configuration between the recorded and "desired" vocalists by a frequency-scale warping of the spectral envelope,

$$\hat{H}(\omega) = H(\omega/\mu), \tag{6}$$

where $H(\omega)$ is the spectral envelope fit to the sinusoidal components in a given frame and μ is a global frequency scaling factor dependent on the pitch modification factor. The factor μ

typically lies in the range $0.75 < \mu < 1.0$. Values of $\mu > 1.0$ could be used to simulate a more child-like voice, as well. It was found that this frequency warping gives the synthesized bass voice a much more rich-sounding, realistic character, and avoids the buzzy artifact.

5 Summary

The LYRICOS system is capable of producing a very natural-sounding, musically pleasing, synthetic singing voice. The LYRICOS system is novel in that it uses "data-driven" methods to model the phonetic information in the voice, resulting in an output that assumes the voice identity characteristics of a recorded human vocalist, and employs a high-quality sinusoidal synthesis method. Furthermore, it is capable of incorporating a wide palette of interesting dynamic expression effects into the output voice signal. A graphical input device based on an industry-standard musical instrument control language provides a mechanism for easy and intuitive manipulation of synthesis parameters by the user.

References

[1] G. Bennett and X. Rodet, "Synthesis of the singing voice," in *Current Directions in Computer Music Research* (M. V. Mathews and J. R. Pierce, eds.), pp. 19–44, MIT Press, 1989.

[2] R. Maher and J. Beauchamp, "An investigation of vocal vibrato for synthesis," *Applied Acoustics*, vol. 30, pp. 219–245, 1990.

[3] J. M. Chowning, "Computer synthesis of the singing voice," in *Sound Generation in Winds, Strings, Computers* (J. Sundberg, ed.), pp. 4–13, Stockholm: Royal Swedish Academy of Music, 1980.

[4] J. M. Chowning, "Frequency modulation synthesis of the singing voice," in *Current Directions in Computer Music Research* (M. V. Mathews and J. R. Pierce, eds.), pp. 57–64, MIT Press, 1989.

[5] P. R. Cook, *Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing*. PhD thesis, Stanford University Department of Music, Stanford, CA, December 1990. Technical Report Stan-M-68.

[6] P. R. Cook, "Synthesis of the singing voice using a physically parameterized model of the human vocal tract," Tech. Rep. Stan-M-57, Stanford University Department of Music, August 1989. also published in *Proceedings of the International Computer Music Conference*, Ohio, 1989.

[7] P. R. Cook, "SPASM, a real-time vocal tract physical model controller and Singer, the companion software synthesis system," *Computer Music Journal*, vol. 17, pp. 30–43, Spring 1993.

[8] K. Lomax, "The development of a singing synthesizer," in *Speech and Computers (SPECOM)*, 1996.

[9] M. W. Macon, L. Jensen-Link, J. Oliverio, M. A. Clements, and E. B. George, "A singing voice synthesis system based on sinusoidal modeling," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 435–438, May 1997.

[10] E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, December 1990.

[11] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 389–406, September 1997.

[12] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, pp. 497–510, March 1992.

[13] E. R. Banga and C. García-Mateo, "Shape-invariant pitch-synchronous text-to-speech conversion," in *Proc. of the Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 656–659, May 1995.

[14] M. W. Macon and M. A. Clements, "Speech concatenation and synthesis using an overlap-add sinusoidal model," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 361–364, May 1996.

[15] E. B. George and M. J. T. Smith, "An analysis-by-synthesis approach to sinusoidal modeling applied to the analysis and synthesis of musical tones," *Journal of the Audio Engineering Society*, vol. 40, pp. 497–516, June 1992.

[16] J. P. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English Speech*. Springer-Verlag, 1993.

[17] B. Pompino-Marschall, "On the psychoacoustic nature of the P-center phenomenon," *J. Phonetics*, vol. 17, pp. 175–192, 1989.

[18] P. Barbosa and G. Bailly, "Characterisation of rhythmic patters for text-to-speech synthesis," *Speech Communication*, vol. 15, pp. 127–137, October 1994.

[19] J. Sundberg, "Synthesis of singing by rule," in *Current Directions in Computer Music Research* (M. V. Mathews and J. R. Pierce, eds.), pp. 45–56, MIT Press, 1989.

[20] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, pp. 820–857, February 1990.

[21] H. M. Hanson, *Glottal Characteristics of Female Speakers*. PhD thesis, Harvard University, Cambridge, MA, May 1995.

[22] M. W. Macon and M. A. Clements, "Sinusoidal modeling and modification of unvoiced speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, November 1997.

[23] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey: Prentice-Hall, 1978.

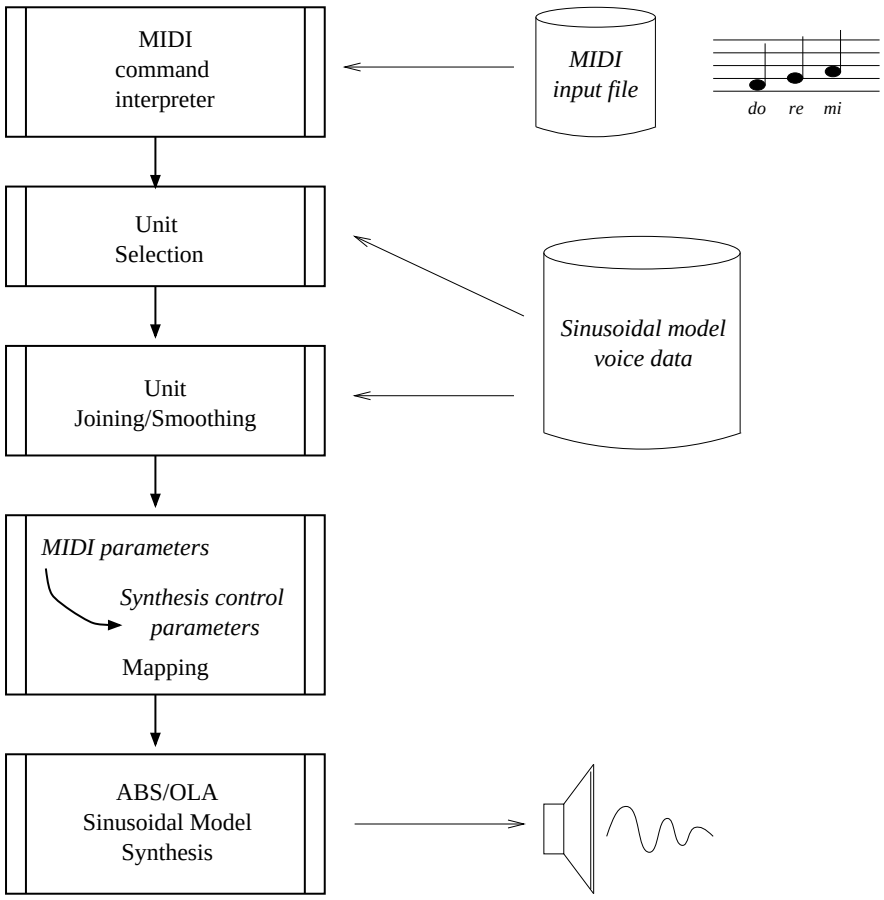

Figure 1: LYRICOS synthesis system block diagram.

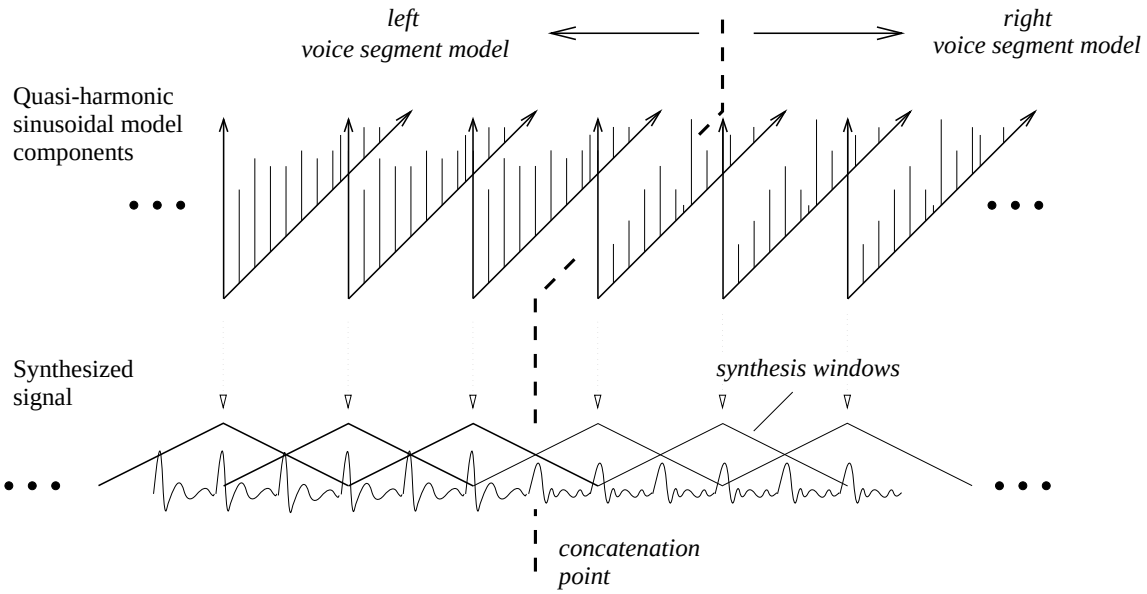

Figure 2: Concatenation of segments using sinusoidal model parameters.

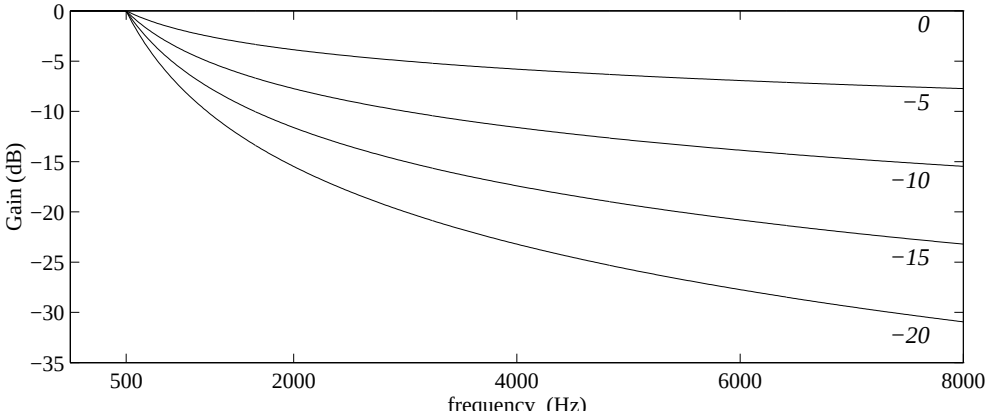

Figure 3: Spectral tilt modification as a function of frequency and parameter T_{in}.