Outlines of Burcas–a Simple Concatenation-based MIDI-to-Singing Voice Synthesis System

Marcus Uneson
Department of Linguistics and Phonetics, Lund University

Abstract

The present paper outlines a simple system (yet to be completed) for concatenation-based singing synthesis in Swedish. The system, called Burcas, takes as input a MIDI file (possibly holding multiple parts) for melody and a text file for lyrics, and it produces standard audio files as output. For the digital signal processing, the MBROLA speech generator is employed.

Burcas consists of an input-parsing interface to three independent phonetic modules: a letter-to-sound converter, a segment duration model, and a data base of sung diphones. In this paper, some considerations of the system as a whole and of its phonetic modules are presented, with emphasis on characteristics of concatenation-based synthesis of singing as opposed to synthesis of speech.

Introduction

This paper introduces Burcas, a (not yet completed) system for singing synthesis that will produce audio files from MIDI[1] data and arbitrary lyrics in Swedish. The system relies on the MBROLA speech generator for all digital signal processing (DSP).

First, a few general issues on the difference between synthesizing speech and synthesizing singing are considered, earlier works on the subject are mentioned, and the MBROLA project is presented. Thereafter, the three phonetic modules of Burcas are briefly commented on: the letter-to-sound (LTS) converter, the segment duration model, and the data base of sung diphones.

Note-to-frequency assignment and other pitch-associated questions are not addressed in this paper, although phonetically relevant. One such non-addressed issue is that of bringing about smooth, portamento-like transitions from one note to another. For a more complete account, see the Burcas web page (Burcas www).

Singing vs speech in concatenative synthesis

The task of synthesizing singing has, of course, a lot in common with that of synthesizing speech. However, there are also important differences—most notably, in singing, the immense problems of reliably modelling intonation and syllable length are already solved (or rather bypassed) by the composer.

For the purposes of a concatenated singing system, at least the following additional properties of singing need consideration:

1) In singing, it is normal for vowels to be sustained

Thus, the segment duration model must be able to handle not only unusually short but also very long syllables. Also sonorant consonants may have very large durations.

2) In singing, the pitch range is wider and higher than in speech

For that reason, the pitch scaling methods must perform well also at high scaling factors, or else the data base(s) must contain various similar tokens recorded at different pitch levels, the most appropriate among which to be chosen during synthesis.

3) In singing, the musical quality of the voice is more critical than the intelligibility of the lyrics

The identity of an individual vowel is often secondary to its intonation and timbre (in particular, female singers, especially sopranos, often modify the lip, tongue, and jaw position so that formants coincide with partials in the source spectrum). Perceived quality differ-

[1]The Musical Instrument Digital Interface (MIDI) standard is an industry-standard protocol for controlling electronic music instruments; it is supported by practically all applications for music, like notation software and sequencers.

ences between vowels tend to disappear anyway as pitch is raised and the spectral envelope of vocal tract resonances is sampled, as it were, more sparsely. As a consequence, the naturalness of vowels is more critical than that of consonants for naturally sounding singing synthesis; however, the exact quality of a vowel is less important. Other phonological contrasts may also be diminished or lost.

The differences mentioned concern in particular the singing style of (educated) opera singers, which is the style that departs most from ordinary speech and also is the most investigated. However, at least the first two items of the list are valid for other genres as well. And all of them are relevant to the phonetic modules of Burcas.

Outlook

Experiments on concatenative synthesis of singing are apparently scarce (although some attempts have been made with other methods, such as formant synthesis). However, a very interesting project for English is known as Lyricos (Macon 1996, Macon et al 1997, Lyricos www). It uses sinusoidal waveform model parameters as concatenation units, selected during synthesis by a specially devised optimizing algorithm. The data inventory is collected from a professional singer. The sinusoidal model permits convenient modification in the frequency domain of spectral properties like spectral tilt. Lyricos takes MIDI as input, as Burcas does. Unlike Burcas, however, Lyricos also interprets MIDI-controllable parameters like vibrato and vocal effort.

The no longer active Lyricos project has a descendant, Flinger, which basically is a customized version of the Festival text-to-speech (TTS) system (Flinger www).

The MBROLA project

The aim of the MBROLA project (MBROLA www) is to boost academic research on speech synthesis by gathering diphone data bases recorded on a voluntary basis for various languages, and providing them freely to the research community (for non-commercial and non-military use). Currently the project offers about 50 data bases in some 25 languages. The data bases must be used with the likewise named MBROLA speech synthesizer, which takes as input a list of phonemes, with their respective duration and frequency, and outputs synthetic, concatenated speech. For the purposes of this project, the MBROLA speech generator is regarded as a black box with three control parameters: phoneme, duration, and pitch.

Overview

Input and output. Interface

As input for melody, Burcas takes a MIDI file, which may contain multiple parts.[2] Lyrics are given in an ordinary text file, with (manually) inserted hyphens for separating syllables and the individual notes of melismatic vowels (as in any sheet of vocal music). Notes and text are aligned syllable-wise. Each voice outputs an audio file in standard format (*.wav, *.au, or *.aiff).

Letter-to-sound conversion

The daunting task of transducing a given input text in the normal orthography of a language into a corresponding phoneme sequence is well-known from TTS systems, and, for most languages, nowhere near to be solved at a more general level.

Each orthographic system provides its own set of difficulties for a TTS system. For Swedish, they include absence of orthographical markings of lexical stress, of morpheme boundaries in compounds, and of word accent; loan words retaining their original spelling (*bourgogne, chianti, rave, aficionado, nachspiel*); various pronunciations of especially the <o> grapheme (*kosta, hosta* ‘cost’, ‘cough’). Of course, such problems recur in LTS converters of many languages, as do those of correctly handling numbers, dates, abbreviations, special characters, etc.

For a singing synthesis system, the problem is different and generally easier. Abbreviations, numbers, etc seldom occur in song texts; if they do, they can always be spelled out. The hyphenation between syllables, mandatory at least in Burcas, in fact bypasses the compound boundary marking problem. The texts are generally short, and the time needed for preparation of the lyrics is probably little compared to the time invested in the music itself. It is therefore of little importance whether or not the

[2]The author wishes to thank Günther Nagler for helpful converters.

LTS converter produces a perfect phoneme transcription at the first attempt. In addition, occasional errors on a segmental level are probably less critical than in speech synthesis, as long as melody is retained.

Two requirements should be fulfilled, though: any graphotactically acceptable input string, although containing unknown words, must produce a valid phoneme sequence with the correct number of syllables; and any corrections must be easy to make.

Given these specifications, a simple rule-based LTS converter will be used for Burcas, basically formalizing well-known reading rules and using a small dictionary of exceptions. A mixture of phonetic symbols and normal orthography will be allowed as input, to facilitate corrections. It is yet to be completed.

Segment duration modelling

The syllable can be regarded as a ”quantal unit” of rhythm, in speech as well as in vocal music. In singing, each syllable of lyric is associated with a number of notes of the melody—one, in syllabic singing, or several, in melismatic.

The rhythmic information of a MIDI file is no more than a set of ”note-on/note-off”- instructions, with associated timepoints. For syllable-timepoint alignment, any given timepoint must be anchored to a specific location in the syllable. An appropriate anchor is the CV-border between onset and nucleus—perceptual experiments have shown that listeners reliably place the beat of a syllable, its ”perceptual center”, at that point (Macon et al 1997).

The segment duration model of Burcas (a modified version of the model presented in Macon et al 1997) is straightforward. For a given syllable, the duration of its associated note or notes is divided between the nucleus of the syllable, the coda of the syllable, and the onset of the subsequent syllable. Each segment has a tabulated value for minimum and ’ordinary’ duration. For each syllable, a scaling factor ρ is calculated:

$$\rho = \frac{\sum_{i=1}^{N_n} Li - \sum_{j=1}^{N_{ph}} D_{\min} j}{\sum_{k=1}^{N_{ph}} (D_{\text{ord}} k - D_{\min} k)},$$

where N_{ph} is the number of phonemes of the syllable, D_{min} and D_{ord} their minimum and ordinary duration, respectively, and N_n the number of notes with duration L associated with the syllable.

If $\rho > 1$, that is, if the syllable is sustained, the nucleus is prolonged. The MBROLA speech generator is not meant for singing and cannot handle very long segments; however, this deficiency can be worked around by concatenating several tokens of the vowel, each of 200-300 ms. (Splitting a long vowel into several, incidentally, also provides a convenient way of emulating a LFO for less static timbre, by letting the duration and/or frequency input of each vowel vary quasi-periodically. This feature is not yet implemented.)

If $0 \leq \rho \leq 1$, the duration D of each segment is calculated as $D_{min} + \rho(D_{ord} - D_{min})$. A negative value of ρ raises an error.

The model, although simple, allows for different compression rates of, say, nasals (whose durations ordinarily are 100-120 ms, but in fast singing may be 50-60 ms) versus plosives (which also typically are 100-130 ms, but seldom shorter than 90-100 ms). More extensive duration data of singing are under preparation.

A more sophisticated approach would of course be to consider quantity as well, absent in the model above. That would make the task of the LTS considerably more difficult and is currently not an issue. It should be noted that the quantity opposition is not always present in singing; for a long syllable (of which there are plenty in singing), the complementary V:C vs VC: syllable structure of central Swedish is obviously difficult to retain. Apparently, the distinction may be lost in more speech-like tempos as well.[3]

The diphone database

The diphone data base currently used (“Ofelia”; female speaker, south Swedish dialect) was produced from and primarily for spoken language. Although it works tolerably as a testing tool, it does have some drawbacks. The timbre is of course not very much like singing. More critical is that the [æ] and [œ] allophones

[3]For instance, duration measurements of lament singing in Estonian, where disyllabic words exhibit a three-way quantity opposition, have shown that the acoustic correlates of quantity present in spoken language largely are lost in song (Ross & Lehiste 1994).

have no transitions but /r/; they thus cannot be employed for sustained vowels.

Recording of a sung data base (male non-trained singer, bass, central Swedish dialect) is scheduled to May, 2002. A possible future extension of the system is the recording of data bases at other pitches as well (preferably by different people); in that way, undesirably high pitch-scaling factors could be avoided. Clearly, a good start is to include a soprano, an alto, and a tenor.

Aims, and not aims

Timbre—or the lack of it

Burcas is a small, zero-budget project. Many interesting questions are not addressed at present—more specifically, anything that cannot be controlled by the three parameters mentioned is silently ignored, as are any MIDI parameters relating to spectrum or intensity.

In particular, it should be stressed that timbre is not an issue. Sampled sounds, when looped, are often perceived as lifeless and rigid by human listeners, and this is particularly true for sampled voices. For one thing, they lack the natural pitch fluctuations that are typical of the human voice; for another, they have nothing corresponding to the increase of vocal effort which a listener connects with singing louder or at a higher pitch (one acoustic correlate of which is a decreased downward tilt of the vocal spectrum). The simplest forms of concatenative methods are very rigid; a lot of low-level difficulties are worked around by handling sampled, predesigned building-blocks. Once a data base is prepared, little can be done to control the spectrum of a given phoneme independently of the fundamental. Quasi-random pitch fluctuations can be emulated with a simple low-frequency oscillator (LFO) mechanism, but characteristics like spectral tilt cannot be controlled. In the long run, the more sophisticated approach of Lyricos approach is certainly worth considering.

Even disregarding signal processing issues, the timbre of Burcas will be rather unexpressive for another reason. The planned diphone data base will be built on recordings not of a trained singer, but of an ordinary choir-member. Where the former usually aims at a personal and powerful voice, the latter rather strives for anonymity.

Possible applications

Burcas does not imitate the performance of a trained singer, nor is it meant to. Given the current limitations, the timbre of the produced singing will be uninteresting at best. However, even so, it might eventually be of use as a research tool, e. g. for studying temporal aspects of phrasing in different musical genres (for instance jazz, ethnic music, lullaby, rap). For such purposes, a non-specialized voice quality might be preferable to a more bel-canto style of singing. Also, such a timbre is better suited for multi-voiced pieces (and, apparently, adding voices in similar rhythms does a lot for the perceived naturalness—probably just by drawing the listener's attention away from the static spectral characteristics).

With a (much) more developed interface, Burcas might in a distant future prove handy as a tool for arrangers of vocal music, in that it can offer synthesized versions of draft arrangements with arbitrary lyrics.

Final remark

This paper has presented the outlines of Burcas, a MIDI-to-singing synthesis system for Swedish employing the MBROLA speech generator. At the time of writing, the system has not advanced very far, and it cannot yet be evaluated. However, any progress will be reported on the project's web page (Burcas www).

References

[Internet resources as for March 2002]

Burcas www
http://www.ling.lu.se/persons/Marcusu/music/burcas/index.html

Flinger www
http://cslu.cse.ogi.edu/tts/flinger/

Lyricos www
http://cslu.cse.ogi.edu/tts/research/sing/sing.html

Macon M (1996). *Speech Synthesis Based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology.

Macon M, Jensen-Link L, Oliverio J, Clements M & George E B (1997). Concatenation-based MIDI-to-singing voice synthesis. In: *103rd Meeting of the Audio Engineering Society*, New York.

MBROLA www
http://tcts.fpms.ac.be/synthesis/mbrola.html

Ross J & Lehiste I (1994). Lost prosodic oppositions: A study of contrastive duration in Estonian funeral laments. *Language and Speech* 37, 407-424.