Articulatory Synthesis of Singing

Peter Birkholz

Institute for Computer Science, University of Rostock
Albert-Einstein-Str. 21, 18059 Rostock, Germany
piet@informatik.uni-rostock.de

Abstract

A system for the synthesis of singing on the basis of an articulatory speech synthesizer is presented. To enable the synthesis of singing, the speech synthesizer was extended in many respects. Most importantly, a rule-based transformation of a musical score into a gestural score for articulatory gestures was developed. Furthermore, a pitch-dependent articulation of vowels was implemented. The results of these extensions are demonstrated by the synthesis of the canon “Dona nobis pacem”. The two voices in the canon were generated with the same underlying articulatory models and the same musical score, the only difference being that their pitches differ by one octave.

Index Terms: Articulatory singing synthesis

1. Introduction

The presented singing synthesizer is based on an articulatory speech synthesizer being developed at our institute since 2001. For the Singing Synthesis Challenge it has been extended to import and process musical scores with lyrics. The common input to the synthesizer for the production of both speech and singing is a gestural score that represents an utterance as an organized pattern of articulatory gestures. The gestural scores are transformed into the movements of articulatory models of the vocal tract and the glottis. These models are in turn mapped on a branched tube model of the vocal system. A comprehensive physical simulation of the flow and acoustical field in the tube system generates the radiated sound.

In Section 2, the components of the articulatory synthesizer will be briefly presented. In Section 3, we will describe the extensions to the synthesizer for the synthesis of singing. A discussion and conclusions follow in Section 4.

2. The articulatory synthesizer

An illustrative overview of the synthesizer is given in Fig. 1. We will start with a short description of the articulatory models of the vocal tract and the vocal folds and then proceed to the acoustical simulation and the generation of speech movements.

2.1. Articulatory models of the vocal tract and the vocal folds

The vocal tract model is a three-dimensional wire-frame representation of the surfaces of the articulators and the vocal tract walls of a male speaker [1, 2]. The shape and position of all movable structures is a function of 23 parameters. By means of magnetic resonance images (MRI) of sustained speech sounds, parameter values were determined for the replication of all German vowels and consonants [2]. Furthermore, using dynamic MRI data, a dominance model has been created that allows to predict the vocal tract parameters for coarticulated consonants.

The vocal folds in our synthesizer are based on a geometrical model proposed by Titze [3] that has been extended for a mechanism of glottal abduction/adduction [4] and an optional parallel chink between the arytenoids [5]. The model predicts the time-varying glottal area at the lower and upper edge of the vocal folds based on parameters for fundamental frequency, pulmonary pressure, degree of abduction, and the size of the optional parallel chink.

For the acoustical simulation, the vocal tract model is transformed into a tube model composed of short abutting elliptical tube sections that can be represented by means of a discrete area function and a perimeter function. The vocal tract tube is combined with tube sections representing the glottis, the trachea, the nasal cavity and the paranasal sinuses. Together, they form a branched tube model of the entire vocal system, which is illustrated by the area function in the middle of Fig. 1. The tube sections for the nasoal cavity are flipped upside-down, and the paranasal sinuses are represented by circles. The tube shape for the nasal cavity was adopted from Dang and Honda [6, 7].

2.2. Aeroacoustical synthesis

For the simulation of acoustics, the branched tube model of the vocal system is represented by an inhomogeneous transmission line circuit with lumped elements [8, 9, 10]. This transformation is based on the analogy of acoustical and electrical transmission lines as described in detail by Beranek [11] and Flanagan [8]. In this circuit model, each tube section corresponds to a two-port network of the T-type, whose elements are functions of the tube geometry. The time-varying distribution of volume velocity and pressure in the whole network is simulated by means of finite difference equations in the time domain with a sampling rate of 44.1 kHz. Dedicated techniques were implemented for the correct simulation of losses due to friction, sound radiation, and wall vibration, as well as for the generation of noise due to turbulence [10, 12]. Therefore, the simulation supports the generation of speech sounds of all major types, like sonorant sounds, fricatives and plosives.

2.3. Generation of speech movements

Speech movements, i.e., the time-varying functions of the parameters of the models for the vocal tract and the vocal folds, are generated on the basis of a gestural score. A gestural score specifies an utterance as an organized pattern of articulatory gestures. This concept is similar to the approaches by Browman and Goldstein [13] and Kröger [14]. However, the definition and execution of the gestures in our synthesizer differs from both former concepts. A detailed description of gestural scores and their execution within the framework of our synthesizer is described in [15, 16]. In this section, only a brief overview will be given by means of the the gestural score for the utter-
3. Extensions of the synthesizer for the synthesis of singing

3.1. Rule-based generation of gestural scores

For the synthesis of singing, we have implemented a few extensions to the speech synthesizer. First of all, a simple xml-format was devised in order to specify the song notes and their attributes. For our demonstration song “Dona nobis pacem”, the file looks as follows.

```xml
<song octaveOffset="0">
    <note beatsPerMinute="110" pitch="rest" type="1/2" vibrato="0.5" lyrics="" loudness="1.0" whisper="0"/>
    <note pitch="g3" type="1/8" lyrics="d o:"/>
    <note pitch="d3" type="1/8" lyrics="o:"/>
    <note pitch="h3" type="1/2" lyrics="n a:"/>
    <note pitch="a3" type="1/8" lyrics="n o:"/>
    <note pitch="d3" type="1/8" lyrics="o:"/>
    <note pitch="c4" type="1/2" lyrics="b i: s"/>
    ...
</song>
```

The most important attributes for a note are the pitch (note letter+octave), type (note length) and the lyrics (here in SAMPA notation). Furthermore, attributes can be specified for the overall speed in beats per minute, the vibrato amplitude in semitones, the loudness, and the degree of whisper. When any of these attributes are not specified for a note, they take the value from the last note, for which they were specified. In our demo...
song, all of the optional attributes are set only once for the initial note (which is actually a rest) and are therefore constant throughout the song. For the transformation of a song into a gestural score, a number of rules were implemented.

The attributes for the loudness and the degree of whisper are simply translated in proportional target parameter values for lung pressure and glottal abduction. The lyrics attribute of a note is first partitioned into an onset, a nucleus, and a coda part. When there is only one nucleus vowel, a corresponding vocalic gesture is generated from the beginning to the end of the note. With two nucleus vowels (a diphthong) the first vowel gesture occupies 65% of the note duration and the second vowel 35%, according to Berndtsson [17]. Consonants of the onset and coda are implemented as consonantal gestures and aligned with the beginning or end of the note, respectively. The duration of consonantal gestures depends on whether or not they occur alone or in a cluster. When they stand alone, they are assigned a predefined inherent duration. In a cluster, their duration is reduced to 80% of their inherent duration for 2 consonants and to 70% for 3 consonants. Depending on the consonant, additional gestures are created in order to open the glottis (for voiceless plosives and fricatives) or to lower the velum (for nasals). The temporal coordination of these gestures with the corresponding consonantal gestures was implemented according to simple “phasing rules”, similar to those described by Kröger [14].

3.2. Pitch dependent vocal tract target shapes for vowels

An important extension to the synthesizer for singing synthesis is the implementation of pitch dependent targets for vowels. It is well known that professional singers often apply different vocal tract shapes to sing the same vowel depending on the pitch of the note. Vowels at higher pitches are often sung with a more “open” articulation and a higher larynx position than vowels at low pitches. In this way, the vocal tract formants are tuned with respect to the harmonics of the voice source. For our synthesizer, we created two “extreme” vocal tract shapes for each vowel – one for low-pitch notes (110 Hz and lower) and one for high pitch notes (440 Hz and higher). The vocal tract target shape for vowels to be sung at pitches between 110 Hz and 440 Hz is linearly interpolated between the low-pitch shape and the high-pitch shape. For the low-pitch shapes, we simply adopted the vocal tract shapes adjusted for speech synthesis. The high-pitch shapes were adjusted manually with respect to a good match between formant frequencies and harmonic frequencies. Special care was taken that the first formant turned out high enough to lie in the vicinity of the first harmonic at 440 Hz.

Figure 2 illustrates the results of this adaptation for the vowel /i/. The midsagittal section of the vocal tract in the top left corner of the picture shows the low-pitch shape for /a/ as derived from MRI measurements for that vowel. The next row shows the vocal tract transfer function (magnitude of the complex ratio between the radiated sound pressure and the volume velocity at the glottis) for the low-pitch shape with the harmonics of the 110 Hz voice source. Obviously, the formant structure is well represented by the harmonic spectrum. The next row shows the same transfer function, but the harmonics for a 440 Hz voice source. Here, the first formant is not at all represented by the harmonic line spectrum. Therefore, the resulting vowel would certainly not sound like a good /i/. The spectrum in the bottom row shows the transfer function of the high-pitch version for /i/ corresponding to the vocal tract in the top right corner of the figure. Here, the first formant has been raised by a higher larynx position and a lower tongue such that it is well represented by the first harmonic of the 440 Hz voice source.

3.3. Generation of $F_0$ targets

The pitch and vibrato attributes of a note are transformed into corresponding gestures for the fundamental frequency. Figure 3 shows the pitch targets generated for a few notes of the demo song. The phonetic transcription of the lyrics (in SAMPA notation) together with the nominal note pitches are shown in the upper part of the picture. Vibrato was generated by short successive pitch targets alternating around the ideal pitch value. In addition to vibrato, we implemented $F_0$ deflections right before and after note boundaries due to “overshoot” and “preparation” inspired by Saitou et al. [18]. The degree of deflection was made proportional to the pitch difference between the successive notes. Finally, fine-fluctuations of $F_0$ were added to the resulting contour.

4. Discussion and Conclusions

An articulatory synthesizer for the synthesis of singing has been presented. The input to the system is an xml-file with the specification of the notes with a number of attributes. This input is transformed in to a gestural score by means of a set of rules. From the gestural score the articulatory speech movements are calculated that finally lead to the generation of the radiated sound. Diverse improvements of the system are conceivable. On one hand, more attributes could be added to the notes to control further properties of the synthetic voice. In this case, new rules would have to be found to translate these attributes
into appropriate articulatory gestures. Furthermore, the adjustment of the low-pitch targets and the high-pitch targets for the vowels should be accompanied and guided by a professional singer. Alternatively, articulatory data of a professional singer could be collected and used to adapt the phoneme targets. Also, the transformation of the lyrics into the corresponding gestures for vowels and consonants needs a lot of fine tuning, especially with regard to the phoneme durations. Nevertheless, we were surprised that an acceptable quality of synthetic singing on the basis of our articulatory speech synthesizer could be achieved with relatively moderate extensions.

5. Acknowledgments

This project was funded by grant no. JA 1476/1-1 from the German Research Foundation.

6. References


