Singing Voice Analysis and Synthesis System through Glottal Excited Formant Resonators

Piero Pierucci (1)
Andrea Paladin (2)

(1) Geosound, Speech and Audio Consulting, Lungotevere Artigiani 28, 00153 Rome, Italy
mc7784@mclink.it
(2) IRIS srl, Parco La Selva 151, I-03018 Italy
mc2842@mclink.it, http://aimi.dist.unige.it/IRIS/index.html

Abstract

This paper presents a system for singing voice analysis and synthesis, that makes use of formant resonators to model the vocal tract behaviour. This well known approach is complemented with the introduction of a modeling of the source signal which is based on an inverse filtering method.

Using this source modeling in combination with a cascade of formant resonators produces a synthesized signal that preserves the timbre characteristics and fine details of the original voice, with the advantage of a compact representation. Furthermore a considerable separation among source and vocal tract characteristics could be obtained, which allows for easy pitch and rate modifications.

The synthesis system is suitable for low cost real-time implementation using fixed point DSP hardware. A MARS platform implementation of the synthesis system is outlined in the paper.

1 Introduction

High quality speech synthesis of the singing voice is, to date, a problem only partially solved, as far as musical applications are concerned. A number of techniques have been proposed to obtain realistic timbres while preserving a compact data representation.

Speech specific approaches generally adopt a source filter characterization with different degrees of sophistication, in order to build synthesis systems that could be controlled from a physiological viewpoint[1][2][3].

More recent approaches are based on a perceptual framework, and generally adopt a non speech specific spectral characterization of the main acoustic events occurring as a result of the speech production mechanism [4].

Both approaches address the problem of obtaining a compact and manageable data representation of speech, in order to allow for speed, pitch and other music related transformations of the original signal. Pros and cons of both approaches can be tough to be at two extremes of a tradeoff line ranging from a physiological consistency of the system model parameters to a psycoacoustical meaningfulness of the adopted signal manipulation framework.

In the present paper a novel approach is investigated, based on an explicit source filter characterization (hence the first of the above mentioned approaches), with a source model that adopts some of the significant results of the non speech specific class of methods, both in terms of data reduction capabilities and perceptually oriented framework for the signal manipulation.

The aim is to obtain an analysis and synthesis system that allow high quality signal reconstruction, even after speed, pitch and other manipulations as morphing and voice transformation, using physiologically meaningful controls. Real-time synthesis is also a concern.

The paper is organized as follows: section 2 reports about speech data acquisition and preprocessing. Section 3 describes in some detail the analysis system, while section 4 reports about a real-time synthesis implementation of the system using the new MARS workstation [5]. Section 5 discusses the results and future directions, and draws some conclusions.

2 Speech data acquisition

In order to allow for best inverse filtering and source characterization, as described in the following sections, the speech signal should be carefully treated before and after digitization.

First the signal is recorded in a silent room, using an high quality condenser microphone, with no filtering and direct analog to digital conversion by a DAT machine. Then the signal is digitally transferred to hard disk for processing. Condenser microphones do allow for best phase and frequency response in the range of interest for source characterization. Even best source characterization could be achieved using special airflow mask systems [6], but at the expense of loss of naturalness on the recorded emissions. After digitization the signal is filtered with zero-phase highpass filtering

with 20 Hz cutoff frequency, to remove very low frequency components.

In order to assess the system for different voices and conditions, professional and occasional singers were recorded. For each voice a recording session was conducted including : a) single vowel constant pitch utterances, over selected notes situated on the natural extension of the singer, for each of the 5 vowels "a, e, i, o, u". b) sequences of the 5 vowels at constant pitch, over the same notes as in a). c) single vowel known melodies, sung at a reference base pitch. Professional singers included soprano, tenor and bass timbres.

3 Speech Analysis

Vocal tract parameters are modeled as a cascade of second order resonators, whose frequency and bandwidths should be estimated from natural data.

This choice is motivated from the fact that formant parameters are well suited for performance control, and that estimation of cascaded resonator parameters could be achieved automatically from the speech signal, with a reasonable degree of quality. Source characterization, in principle, could then be obtained in a reliable way when vocal tract resonance are removed from the speech signal, through a method often called "inverse filtering".

The above scheme is only partially true if high quality synthesis is the task, mainly because of source tract interaction phenomena[7], that make it difficult to estimate both second order resonator characteristics and the source signal, especially for high pitched (female) voices. The best approach would be the joint estimation of source and filter parameters, which unfortunately is a non linear problem. Sub optimal techniques are then adopted to achieve good source and vocal tract separation during joint estimation, and post-processing is applied to smooth and improve this estimation. At the end of this stage, vocal tract parameters are said to be estimated. This corresponds to a fixed displacement framed analysis which results in a sequence of vectors related to the resonator parameters and source data representing the residual signal obtained from inverse filtering with formant data. Then source parameters are determined from raw estimated data to obtain substantial reduction in the data rate.

In the following, details and examples of the above procedure are given, separately for vocal tract and source estimation parts.

3.1 Vocal tract characterization

Vocal tract model parameters are estimated through an iterative method, sketched in figure 1, based on inverse filtering, similar to IAIF (Iterative Adaptive Inverse Filtering)[8], in which the LPC analysis steps are carried out using the RLPC (Robust Linear Prediction) approach [9].

The use of LPC based criteria to determine formant resonance parameters (frequencies, bandwidths and aptitudes) is well known to give satisfactory results when source and vocal tract do not interact too much. Generally for female (and singing) voices, worse performances are to be expected, especially for the first formant, due to the presence of residual oscillations from the previous pitch period that could interact with actual excitation in a destructive or constructive manner. To this extent the use of RLPC is beneficial to reduce the bias of the estimation. This technique applies iteratively LPC estimation to the weighted input signal. The weighting function is calculated, at each iteration, to gives less weight to covariance matrix entries that contribute more to the LPC residual error.

Furthermore the non ideal (non pulse-like) behaviour of the source signal contributes to the bias of the spectral estimator, and the use of IAIF is beneficial in order to feed the RLPC estimator with an input signal from which the glottal contribution to the overall spectrum tilt has been removed.

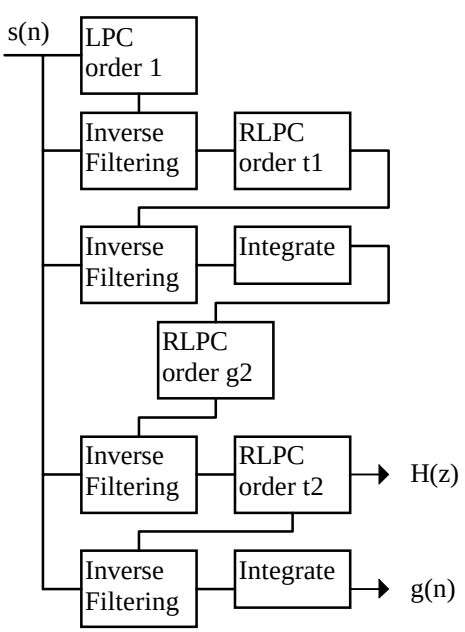

Figure 1 Block diagram of the source-filter estimation procedure: s(n) : input signal. H(z) : vocal tract only transfer function. g(n) : source signal.

After determination of the vocal tract response H(z), formant resonator parameters are determined through root solving of the LPC polynomial.

In the case of female soprano singing at very high pitch, the above method still gives results that should be post-processed to obtain smooth trajectories for formant frequencies and bandwidths. Due to the method

adopted, anyway, simple median filtering schemes have proven to be sufficient to consistently smooth the trajectories. In figure 2 a plot of first 6 formants and bandwidths of a female soprano singing the vowel sequence "aeiou" is reported. Vibrato effects are seen to influence the results of formant tracking.

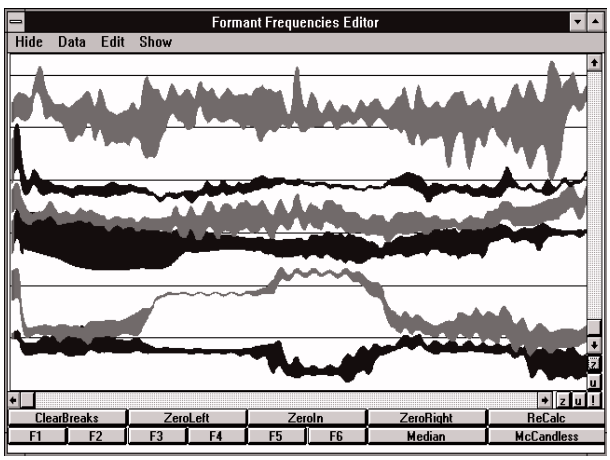

Figure 2 : Formant (and bandwidth) tracks of a professional soprano voice, singing the vocalic sequence "aeiou". *x* timescale: 10 sec. *y* frequency scale: 6.0 kHz. Average pitch is 360 Hz.

The proposed method compares favorably with pitch synchronous closed phase formant tracking [10], in that it does not need to locate glottal closure instants.

Finally, after trajectories smoothing, filtering the original signal through a cascade of 2nd order inverse filters and an inverse lip radiation model [6] is performed, to obtain the glottal signal input for the source characterization stage.

3.2 Source characterization

Once obtained an estimation of the glottal source signal from the previous step, a pitch synchronous analysis is carried out for each frame to extract a period of the source signal to be spectrally manipulated and temporally interpolated. Then a subset of those glottal samples is retained to be used at synthesis time, and cosinusoidal temporal interpolation curves between glottal samples are also derived.

The extraction of glottal periods is obtained through suitable time windowing 2 pitch periods of glottal signal [11] around the center of the analysis frame.

Spectral manipulation include DFT, zero padding , zero phasing, and inverse FFT, to provide a fixed length interpolated source signal to be used as wavetable for synthesis. Zero phasing allows for easy temporal interpolation schemes, suitable for real-time operation.

A subset only of the glottal periods is retained for synthesis. The selection is carried out following criteria based on normalized spectral harmonic distance between glottal periods [12].

The choice to not explicitly characterize the source signal with parametric models, as in [3], is based on real-time synthesis constraints, and on the belief that this approach would balance limitations of the currently adopted vocal tract model.

4 Synthesis System

The synthesis structure has be implemented both on software and on a MARS workstation[5], to evaluate the performance degradation due to the 24 bit fixed point arithmetic used in the real time synthesis algorithm on the MARS. The fixed point synthesis scheme seems to be acceptable up to a cascade of 8 resonant filters. Multiple delay resonant filters has been adopted to extend the bandwidth of the synthetic signal while using a reduced number of formants.

All the analysis data are loaded on the MARS sample memory before synthesis. Then they are used by the real-time synthesis system with no intervention of the host computer.

The synthesis algorithm is divided into two main blocks: the glottal source oscillator and the cascade resonant filters. The first block is composed by two bandwidth controlled oscillators reading consecutive pre-stored wavetables. These oscillators are cross-faded using suitable weighting functions to avoid clicks during transitions. Then a gated and filtered noise is added to the oscillators output to simulate the noisy excitation component typical of breathy vocal emission. Pitch can be reconstructed from stored original contours or assigned to external controls.

The second block is composed by six, or more, cascaded multiple delay resonant filters, whose coefficients are linearly interpolated from values stored in MARS sample memory.

The main advantage of using MARS is that the algorithms can be developed and debugged in a graphical and interactive environment.

The experimental setup takes advantage of the MARS system capabilities, which include up to 8 independent outputs. One of the outputs allows to listen to the glottal excitation reconstruction. Three more outputs are used for the original signal, the synthetic signal and the floating point synthetic signal, as produced by the off line simulation system.

Using this source/filter approach, pitch and rate manipulations are straightforward. To control the proposed algorithm in a MIDI environment, a set of timbres have been defined with the following relationship between MIDI controls and algorithm parameters:

Param.Name	Midi Ctrl	Comment
ext_int_freq	Expression	switch original pitch/keyb.
scale_freq	Modulation	scale original pitch
ext_freq	Key+PB	keyboard pitch & bending
soglia_trig	Pan	noise gate duty cycle
amp_noise	Breath	noise amplitude
hicut_noise	Foot	noise hipass cutoff freq.
locut_noise	Portamento	noise lowpass cutoff freq.
band_pulse	Balance	wavetable stretch percent
syn_speed	Volume	time stretching

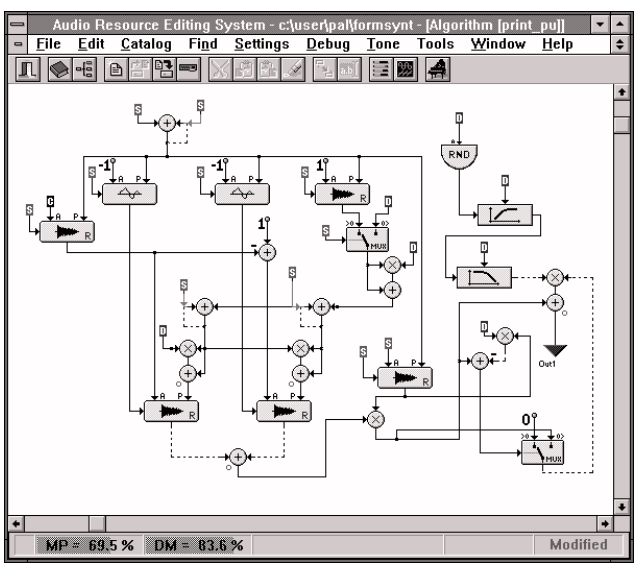

Figure 3: Glottal Source Model view through the ARES algorithm editor.

5 Conclusions

In this paper a system for singing voice analysis and synthesis, that make use of formant resonators to model the vocal tract resonances, and glottal based wavetable synthesis to represent the source signal, has been presented. The system is suitable for real-time operation.

Informal listening tests indicate that the method enables one to reconstruct the original timbre of the singer with high quality, allowing for time, pitch and timbre manipulation.

Future extensions of the work include a detailed study of suitable glottal periods selection criteria, and a comparison with explicit glottal models for source signal reconstruction.

6 Acknowledgments

The authors would like to thanks the MARS/ARES development team at IRIS, for the powerful tools which allowed the completion of this work. We also would like to thank Mr. Giuseppe Di Giugno, for his support and helpful discussions.

References

[1] Cook, P.R., 1989, “Synthesis of the Singing Voice Using a Physically Parametrized Model of Human Vocal Tract”, proc. of ICMC89, pp.69-72

[2] Pinto, N.B., and Childers, D.G., 1989, “Formant Speech Synthesis: Improving Production Quality”, IEEE Trans. Acoust. Speech Signal Process., ASSP-37, 1870-1887.

[3] Milenkovich, P.H., 1993, “Voice Source Model for Continuous Control of Pitch Period”, Journal of Acoustic Society of America, Volume 93, Number 2, pp.1087-1096.

[4] George, E.B., and Smith, M.J.T., 1992, “Analysis by Synthesis/Overlap-Add Sinusoidal Modeling Applied to the Analysis and Synthesis of Musical Tones”, Journal of Audio Engineering Society, Volume 40, Number 6, pp.497-515.

[5] Andrenacci, P. et al., 1997. “The new MARS Workstation” , ICMC97, this issue.

[6] Javkin, H.R. et al., 1987. “Digital Inverse Filering for Linguistic Research”, Journal of Speech and Hearing Research, Volume 30, pp.122-129.

[7] Fant, G., 1993. “Some problems in voice source analysis,” Speech Communication, Volume 13, pp. 7-22.

[8] Alku, P., 1992, “Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering”, Speech Communication, Volume 11, pp.109-118.

[9] Lee, C.H., “On Robust Linear Prediction of Speech”, 1988, IEEE Transactions on Acoustics, Speech and Signal Processing, Volume 36, Nuber 5, pp.642-650.

[10] Pierucci, P., Mesirca, A., 1995, “Automatic Determination of Singing Voice Formants”, Proc. of CIARM95, Ferrara, Italy, pp.151-156.

[11] Bristow-Johnson, R., 1995, “A Detailed Analysis of a Time-Domain Formant-Corrected Pitch-Shifting Algorithm”, Journal of Audio Engineering Society, Volume 43, Number 5, p.348.

[12] Horner A., Beauchamp, J., 1996, “Piecewise Linear Approximation of Additive Synthesis Envelopes: A Comparison of Various Methods”, Computer Music Journal, Volume 20, Number 2, pp.72-95.