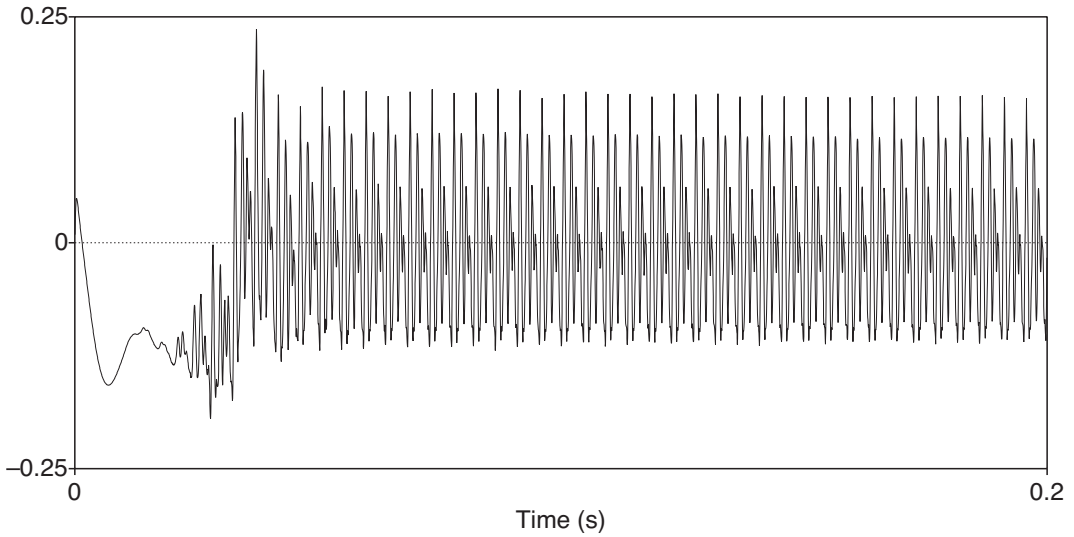
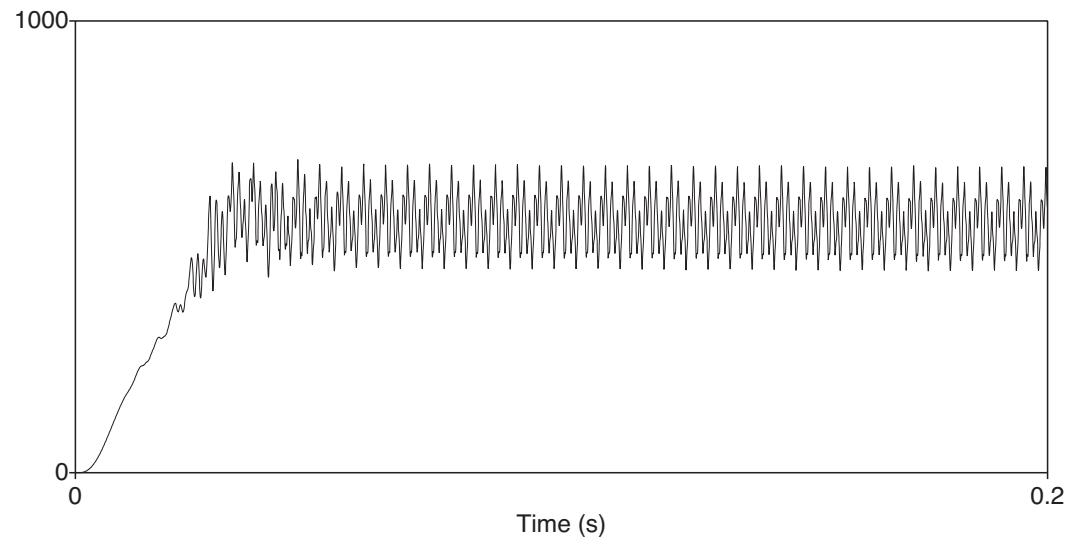
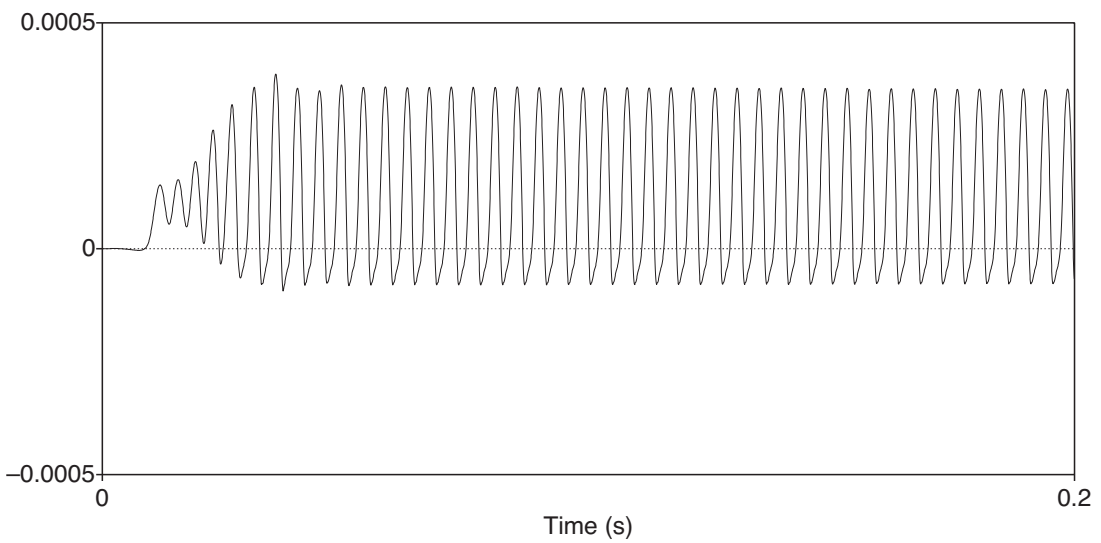






Frontispiece. (a) The first 0.2 seconds of the synthesized open vowel sound. (b) The excitation pressure at the bottom region of the trachea section of the model. (c) The behavior of the vocal folds of the model as a function of the width of the glottis during vibration. (© Eduardo Reck Miranda)

Artificial Phonology: Disembodied Humanoid Voice for Composing Music with Surreal Languages

Eduardo Reck Miranda

ABSTRACT

This paper examines some of the core techniques used to create an artificial phonological system for *Sacra Conversazione*, a short opera in five acts featuring human singers, artificially synthesized voices and complementary electroacoustic sounds. It introduces some of the most significant techniques for computer simulation and manipulation of voice used to produce materials for the piece, namely physical modeling, additive re-synthesis, PROSE and PSOLA. The author concludes with a discussion of lessons learned throughout the process of composing with such techniques.

The main motivation for the composition of *Sacra Conversazione* was to explore the possibility of composing with the sounds of a surreal pseudo-language, in a primarily linguistic sense rather than a musical sense. The piece is entirely sung with "words" that do not exist in any language that I am aware of, forming what I refer to as the utterances of a surreal pseudo-language. The creation of these artificial utterances was informed by studies of the phonetics of various languages [1] and experiments in combining recorded syllables from these languages: Amharic, Catalan, Cantonese, Croatian, Dutch, Galician, German, Hebrew, Hindi, Irish, Persian, Swedish, Thai and Turkish. The lyrics (Fig. 1) were almost entirely written on the score using the phonetic notation proposed by the International Phonetics Association (IPA) [2]. A variety of nonstandard vocals and onomatopoeic vocalizations also have been produced by programming the computer to generate surreal utterances that humans would find extremely difficult, if not impossible, to produce.

SIMULATING THE VOCAL SYSTEM

The vocal system can be thought of as a resonating structure in the form of a complex pipe. The acoustic properties of a straight, regular pipe are well understood and fairly straightforward to model, but the shape of the vocal tract is neither straight nor regular. On the contrary, it is bent, its length can be stretched and contracted, its diameter is variable at different sections and it branches into two main paths: one to the mouth and the other to the nasal cavity. Length, diameter and branching control for producing nasal sounds change during sound production.

Eduardo Reck Miranda (composer, research scientist), Faculty of Technology, University of Plymouth, Plymouth, PL4 8AA, U.K. E-mail: <eduardo.miranda@plymouth.ac.uk>. Web site: <neuromusic.soc.plymouth.ac.uk>.

Sound samples related to this article are available at: <cmr.soc.plymouth.ac.uk/lmj15_examples/>.

The vocal system can be roughly simulated as a subtractive synthesizer consisting of a source module and a resonator module. The source module produces a raw signal intended to simulate the waveform produced by the vibration of the vocal folds, which in turn is shaped by the acoustic response of the resonator module. The resonator is implemented as an arrangement of band-pass filters, each of which is tuned to resonate at the frequencies of the most prominent standing waves of the vocal tract. These standing waves correspond to the distinct resonance that confers the characteristic timbre of the human voice [3]. Normally five band-pass filters arranged in parallel are sufficient to produce realistic vocal sounds [4–6].

Physical Modeling: Praat

The physical model used by a system called Praat, implemented at the Institute of Phonetic Sciences, University of Amsterdam, is a more sophisticated method for simulating the vocal tract than subtractive synthesis [7].

In this model, the whole vocal system is represented as a squared pipe composed of dozens of shorter pipes, whose walls are represented by mass–spring-damping (MSD) units. Roughly speaking, the springs in Fig. 2 represent muscles in the sense that their rest positions and tensions can be altered in order to adjust the walls and the internal obstructions of the pipe. The central idea here is that the walls and obstructions yield to air-pressure changes, and air is forced to flow inside the pipe as the result of mass inertia and elasticity. For example, by diminishing the volume of the lungs, one generates displacement of air towards the mouth; in other words, excitation causes phonation.

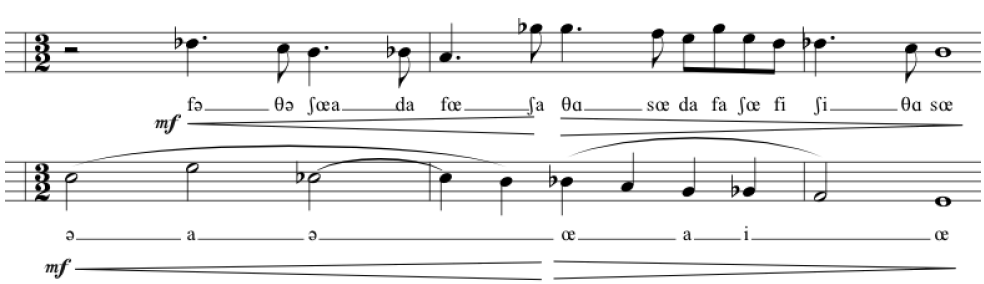


Fig. 1. ***Sacra Conversazione*, composed between 2000 and 2003. (© Eduardo Reck Miranda) An excerpt of the score with the lyrics for a soprano written using phonetic notation.**

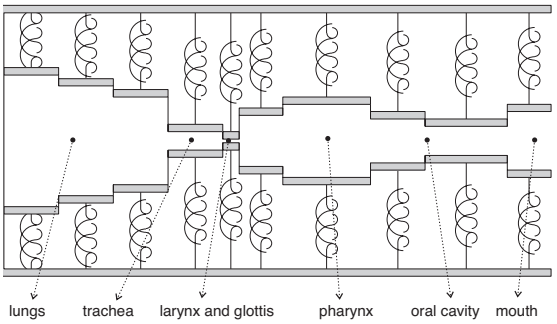


Fig. 2. The vocal system can be simulated with a number of linked short pipes whose walls are represented by mass–spring-damping units. (© Eduardo Reck Miranda)

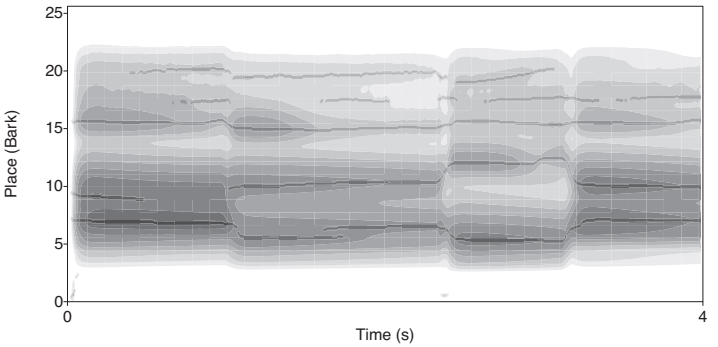


Fig. 3. The cochleogram of a sequence of four vowels. (© Eduardo Reck Miranda)

The damping element is not shown in Fig. 2, but each spring of the model does have a damping element associated with it. Two types of equations are needed to compute the simulation: myoelastic equations that describe the physical behavior of the walls of the pipes and aerodynamic equations that describe the evolution of the movement and pressure of the air in the network. The variables for these equations are specified by means of 29 parameters that metaphorically describe the actions of the vocal tract muscles and organs, such as the cricothyroid, styloglossus, orbicularis oris and masseter.

Sculpting the Voice: Analysis and Re-Synthesis

In contrast to simulation of the vocal system, the following techniques focus on modeling the sounds themselves and not on the functioning of sound-producing mechanisms. Such techniques are chiefly based upon the notion of analysis and re-synthesis of sound [8].

Analysis and re-synthesis can be performed in a number of ways. During the analysis stage a number of parameters are extracted from a sampled sound. The re-synthesis stage then uses these parameters to re-create the sound using a suitable synthesis technique such as additive re-synthesis.

Short-Time Fourier Transform (STFT) is an adaptation, suitable for computer programming, of the original Fourier-analysis mathematics for calculating harmonic spectra. STFT chops the sound into short segments called windows and analyzes the spectrum of each segment sequentially. Each window of the analysis contains two types of information: a magnitude spectrum depicting the amplitudes of every analyzed component and a phase spectrum showing the initial phase for every frequency component.

Additive Re-Synthesis: Diphone

Additive re-synthesis employs STFT analysis data to control an additive synthesizer [9]. In this case, the algorithm converts STFT data into amplitude and frequency trajectories (or envelopes) that span the width of the STFT frames. As envelope data are generally straightforward to manipulate, transformations such as stretching, shrinking, rescaling and shifting can be applied to either or both frequency and time envelopes prior to the re-synthesis process.

An example of an implementation of additive re-synthesis can be found in Diphone, a program for making transitions between sounds developed at the Institut de Recherche et Coordination Acoustique/Musique (IRCAM) in Paris [10]. This software addresses one of the most challenging synthesis problems of artificial speech using sampled phonemes: the concatenation of sequences. The computer's rendition of a specific syllable, such as /mu/, may not always sound satisfactory. For instance, consider the words *music* and *mutation*. The computer's rendition of the syllable /mu/ in the word *music,* if used at the beginning of the word *mutation*, sounds artificial because the transition between /u/ and /s/ in the word *music* and the transition between /u/ and /t/ in the word *mutation* involve different spectral behaviors. This transition problem is also evident in music composition, in which certain articulations and musical passages are clearly more appropriate to our ears than are others.

Each sound sample in Diphone is subjected to spectral analysis to extract information on how its spectrum evolves in time. This analysis provides a multiparametric representation of the sound; it contains information about its fundamental frequency (i.e. pitch), plus the frequency, the amplitude and the phase of its spectrum components. The advantage of such representation over straight sampling is that this information can be manipulated individually and mapped onto the parameters of a synthesis algorithm

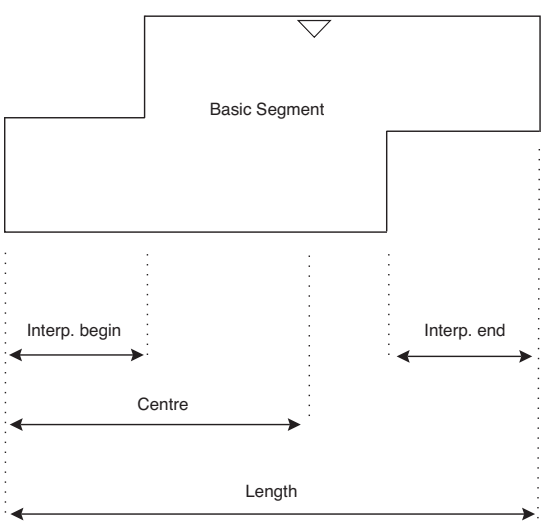
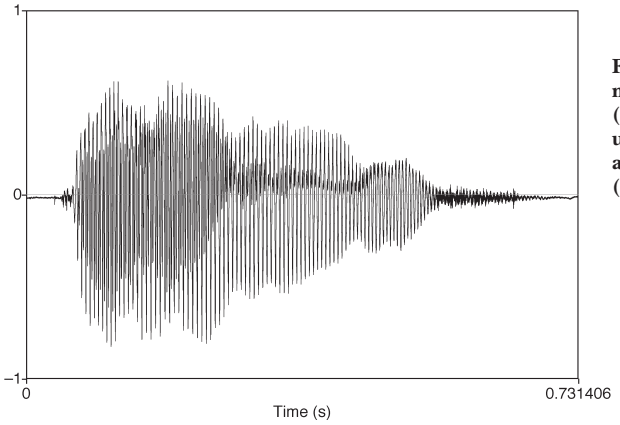


Fig. 4. Each segment in Diphone is represented by an icon with three distinct areas: a central area and two adjacent interpolating areas. (© Eduardo Reck Miranda) The user can adjust the extent of the interpolating areas and define the type of interpolation algorithm for each side.



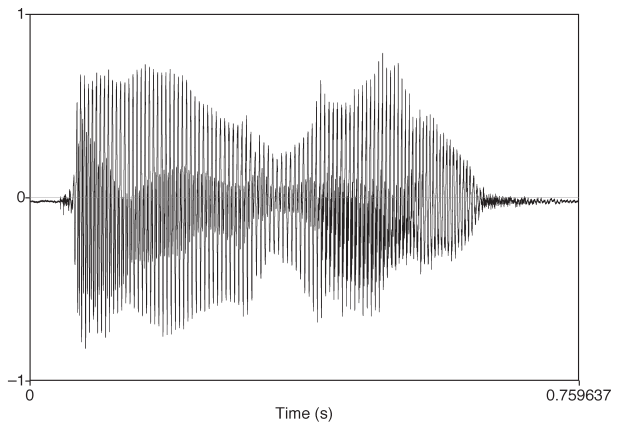
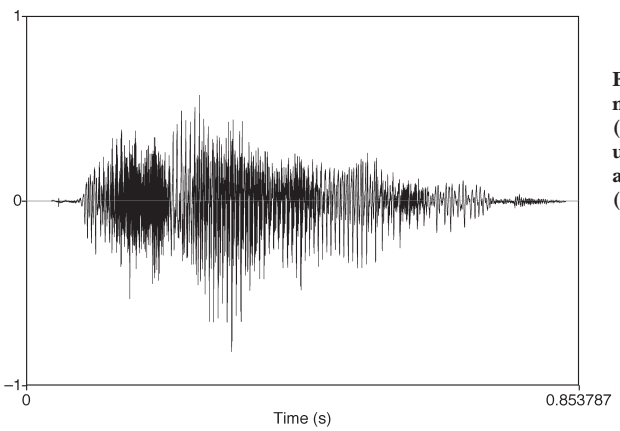


Fig. 5. Two Irish words, meaning (a) "feared" and (b) "yellow" in English, used for the creation of an artificial utterance. (© Eduardo Reck Miranda)



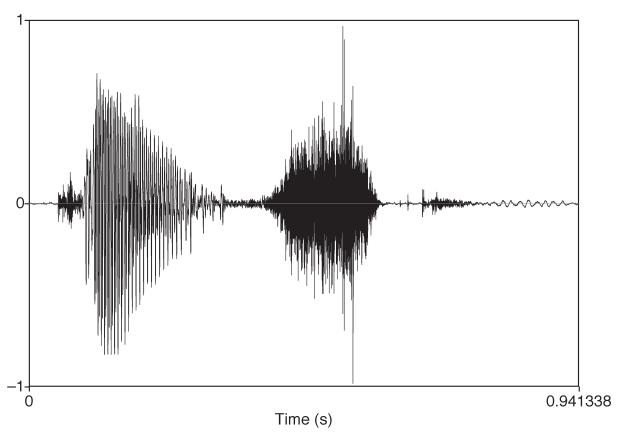


Fig. 6.Two Croatian words, meaning (a) "loan" and (b) "bone" in English, used for the creation of an artificial utterance. (© Eduardo Reck Miranda)

such as additive synthesis. If the analysis information is changed, then the resulting re-synthesis will sound different.

Prosody Extraction: PROSE

Prosody can be defined as the information in human speech that is not conveyed by the sequence of words. Prosody is the music of speech, created by the patterns of stress and intonation in a spoken signal. Prosody Extraction (PROSE) extracts prosodic information from a given spoken sentence and uses this information to re-synthesize an open vowel (e.g. /a/ as in the word "car" in English) with such prosody added [11]. The idea is that the sentence is stripped of a great part of its semantics, as the words are no longer present; only the prosody remains.

PROSE starts by extracting the pitch envelope and the energy contour of a speech stream. Then, the extracted pitch envelope is fed into a subtractive formant synthesizer that produces an open vowel sound with variations in the fundamental frequency driven by the input pitch envelope. Next, the result of the synthesis is modulated by the extracted energy contour. The outcome of this process is the extracted prosody of the input speech signal.

The pitch-extraction module employs an autocorrelation-based technique. Autocorrelation works by comparing a signal with segments of itself delayed by successive intervals, or time lags: starting from one sample, two samples, etc., up to m samples. The objective of this comparison is to find repeating patterns that indicate periodicity in the signal. Part of the signal is held in a buffer, and as more of the same signal flows in, the algorithm tries to match a pattern in the incoming signal with the signal held in the buffer. If the algorithm finds a match (within a given error threshold), then there is periodicity in the signal, and in this case the algorithm measures the time interval between the two patterns to estimate the frequency [12].

The synthesis task is performed by a subtractive synthesizer. Since the objective is to produce an open vowel throughout the utterance, the values for the formant frequencies, formant bandwidths and attenuation coefficient are set to produce an open vowel (e.g. /a/ as in the word "car").

The energy contour is obtained by convolving the squared values of the samples with a smooth bell-shaped curve with very low peak-side lobes (e.g. –92 dB or less).

Finally, the modulation stage employs the amplitude modulation technique

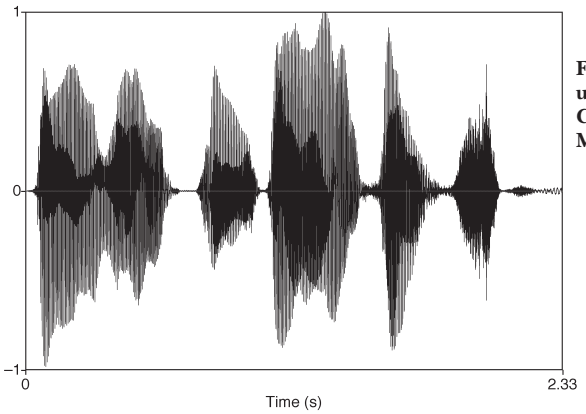


Fig. 7. An artificial utterance using syllables from Irish and Croatian. (© Eduardo Reck Miranda)

whereby the synthesized utterance is multiplied by a normalized version of the energy contour.

Prosody Manipulation: PSOLA

Pitch Synchronous OverLap and Add (PSOLA) works by concatenating small segments with an overlapping factor [13]. The duration of these segments is proportional to the pitch period of the resulting signal. PSOLA is useful because a given signal can be decomposed in terms of these elementary segments and then re-synthesized by streaming these elements sequentially. Parametrical transformation can be applied during the streaming of the segments in order to modify the pitch and/or the duration of the sound. PSOLA is particularly efficient for shifting the pitch of a sound, as it does not, in principle, change its spectral contour.

At the analysis stage, the elementary segments are extracted using windows centered at the *local maxima* (i.e. regular cycles displaying a single energy point), called markers. These markers should be pitch synchronous in the sense that they should be close to the fundamental frequency of the sound. The size of the window is proportional to the local pitch period. The pitch of the signal can be changed during re-synthesis by changing the distances between the successive markers. Time stretching or compression is achieved by repeating or skipping segments.

The caveat regarding the PSOLA analysis and re-synthesis method is that it only works satisfactorily on sounds containing the local maxima. The good news is that human vocal sounds, especially vowels, fulfill this requirement.

Examples

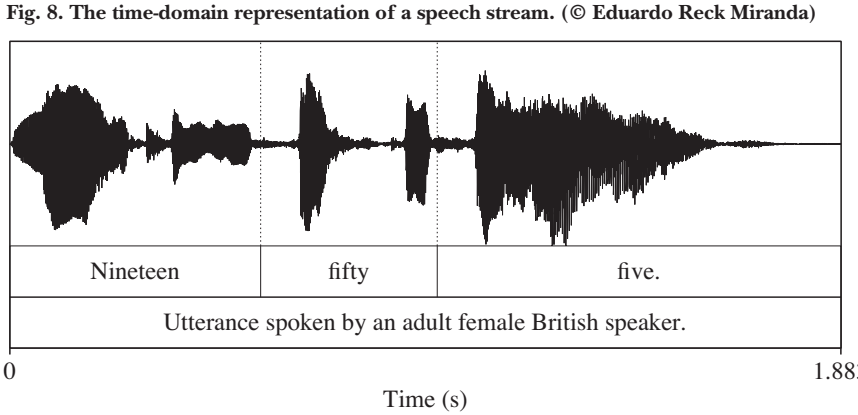
Physical Modeling Synthesis Example

In order to synthesize a sound using Praat's physical model, one needs to set the details of the speaker model and a list of actions for this model. For the sake of simplicity, the following example employs the system's default adult female speaker model, and I will refrain from changing its physical characteristics.

The list of actions for the speaker is given in a script defining breakpoint values for the 29 variables of the model, as mentioned earlier. These variables are highly dependent on one another, and in theory most of them can vary from –1.0 to +1.0. In most cases we can assume that the starting point is at 0.0, which stands for the relaxed position of the corresponding part. The following example involves only seven variables. The remaining 23 are set to the default system values, which normally correspond to the resting positions of the respective parts that they represent. The variables are as follows:

- Interarytenoid (glottis): This variable influences the width of the glottis (i.e. the space between the vocal folds). It normally varies from 0.0 (vocal folds relaxed and open) to 1.0 (vocal folds stiffly closed). A value of 0.5 brings the vocal folds together into a position suitable for normal voicing.
- Cricothyroid (glottis): This variable influences the length and the tension of the vocal folds. It can normally vary from 0.0 to 1.0. Contraction of the cricothyroid muscle in a real vocal system causes a visor-like movement of the thyroid cartilage on the cricoid cartilage; this movement lengthens the vocal folds. The pitch of the utterance rises proportionally to the value of this variable.
- LevatorPalatini (velum): This variable controls the velopharyngeal port to the nasal cavities, where 0.0 opens the port and 1.0 closes it by lifting the velum. An open port (0.0) will nasalize the sound. Nasal sounds have fewer high-frequency components in their spectrum.
- Hyoglossus (tongue): This variable causes the tongue body to be pulled backward and downward. It normally varies from 0.0 (resting position) to 1.0 (maximally pulled).
- Mylohyoid (mouth): This variable influences the opening of the jaw and moves the back of the tongue toward the rear pharyngeal wall. It is normally set with values from 0.0 (resting position) to 1.0 (maximum jaw opening). A negative value (e.g. –0.5) can be used to raise the body of the tongue.
- OrbicularisOris (mouth): This variable bears the name of a muscle that circles the lips. It influences the rounding and protrusion of the lips of the model and normally varies from 0.0 (lips spread normally) to 1.0 (lips rounded with protrusion).
- Lungs: This variable is responsible for producing the airflow that sets the vocal folds into vibration during phonation. This variable normally

Fig. 8. The time-domain representation of a speech stream. (© Eduardo Reck Miranda)



has values between –0.5 and +1.5. The value –0.5 represents the maximum amount of air that can be exhaled by force, and the value +1.5 represents the maximum amount of air that can be inhaled. A typical exhalation values 0.0, and a typical inhalation would not exceed 0.2 for normal voicing.

The following script produces an open vowel:

```
#— — — — — — — — — — — — — —
# Example: A simple phonation
#— — — — — — — — — — — — — —
Create Speaker... Robovox Female 2
Create Artword... phon 0.5
#— — — — — — — — — — — — — —
# Supply lung energy
#— — — — — — — — — — — — — —
Set target... 0.00 0.1 Lungs
Set target... 0.03 0.0 Lungs
Set target... 0.5  0.0 Lungs
#— — — — — — — — — — — — — —
# Control glottis
#— — — — — — — — — — — — — —
# Glottal closure
Set target... 0.0 0.5 Interarytenoid
Set target... 0.5 0.5 Interarytenoid
#
# Adduct vocal folds
Set target... 0.0 1.0 Cricothyroid
Set target... 0.5 1.0 Cricothyroid
#— — — — — — — — — — — — — —
# Close velopharyngeal port
#— — — — — — — — — — — — — —
Set target... 0.0 1.0 LevatorPalatini
Set target... 0.5 1.0 LevatorPalatini
#— — — — — — — — — — — — — —
# Shape mouth to open vowel
#— — — — — — — — — — — — — —
# Lower the jaw
Set target... 0.0  –0.4 Masseter
Set target... 0.5  –0.4 Masseter
#
# Pull tongue backwards
Set target... 0.0  0.4 Hyoglossus
Set target... 0.5  0.4 Hyoglossus
#— — — — — — — — — — — — — —
# Synthesize the sound
#— — — — — — — — — — — — — —
Select Artword phon
plus Speaker Robovox
To Sound... 22050 25 0 0 0 0 0 0 0 0 0
```

The command *Create Speaker* needs three parameters: (1) the name of the speaker, (2) the kind of speaker and (3) the type of the glottis. Our speaker is called Robovox, and it uses the female tract model that is defined by default in Praat. In order to excite its vocal folds, Robovox needs to produce pressure flow by reducing the equilibrium width of the lungs. This is achieved by reducing the value of the *Lungs* variable from 0.1 to 0.0 during the initial 0.03 seconds of the utterance "Supply lung energy." Also, Robovox has to keep its glottis closed with a certain stiffness so that the vocal folds can vibrate with the pressure coming from the lungs. This is achieved by setting the *Interarytenoid* and the *Cricothyroid* variables equal to 0.5 and 1.0, respectively, during the whole utterance "Control glottis." Then, Robovox should close its velopharyngeal port to the nasal cavity in order to prevent air escaping through the nose, otherwise the sound will lack energy in its high-frequency components. Robovox's velopharyngeal port can be closed by setting the *LevatorPalatini* variable to 1.0 during the whole utterance "Close velopharyngeal port." Finally, Robovox should open its mouth and shape the vocal tract in order to produce a vowel. The mouth can be opened by setting the *Masseter* to –0.4. The *Hyoglossus* variable is set to 0.4 in order to pull Robovox's tongue backward and downward; this shapes the mouth to produce an open vowel ("Shape mouth to open vowel").

The next and final step is to activate the synthesis engine to produce the

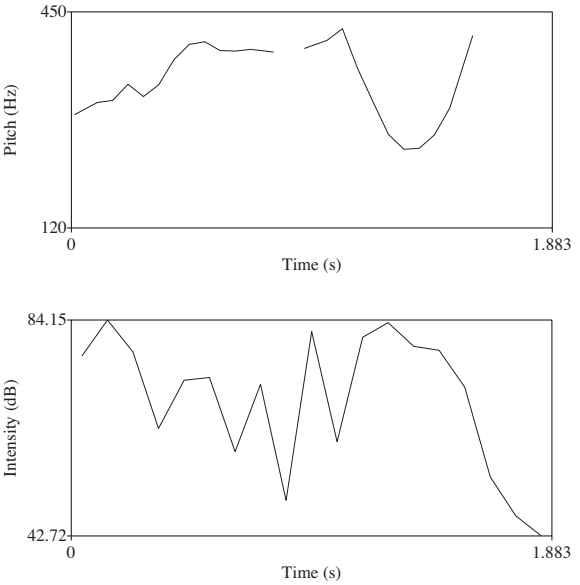
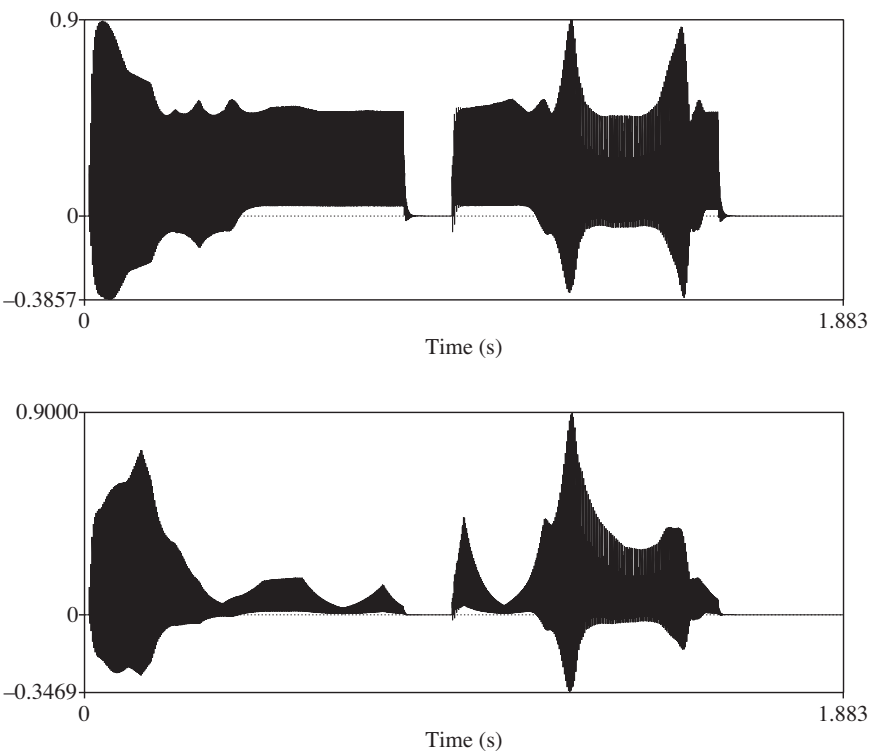


Fig. 9. (a) The extracted pitch envelope of the utterance shown in Fig. 8. (b) The energy contour of the utterance shown in Fig. 8. (© Eduardo Reck Miranda)

Fig. 10. (a) The synthesis result prior to modulation. (b) The outcome of PROSE. (© Eduardo Reck Miranda)



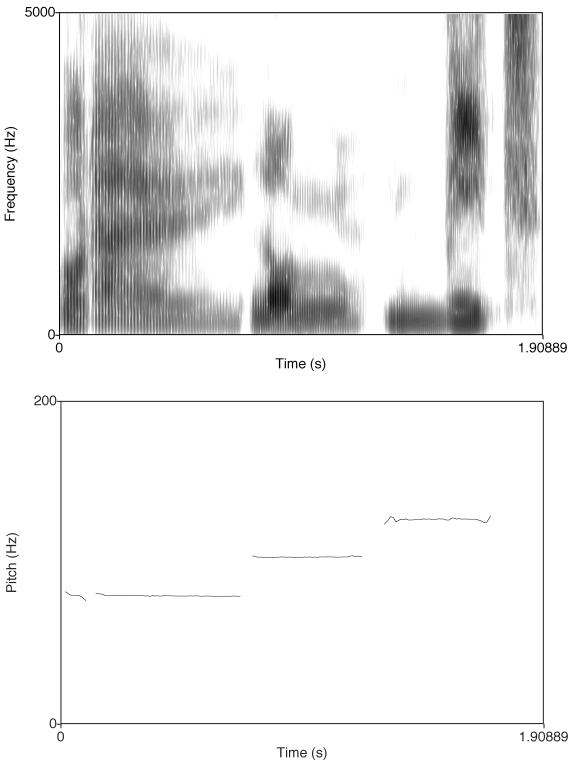


Fig. 11.(a) The spectrogram of an artificial utterance composed of three syllables and (b) its pitch analysis. (© Eduardo Reck Miranda)

sound ("Synthesize the sound"). This is done by activating the synthesizer with the command *To Sound,* whose first parameter is the sampling rate of the sound that will be synthesized (in this case 22,050 Hz) and whose second is the oversampling coefficient. Oversampling is the number of times that the physical modeling equations will be computed during each sample period; the default value is 25. The two first lines of this section say that we wish to synthesize the sound *phon* using the speaker Robovox.

The top graph shown on the Frontispiece portrays the first 0.2 seconds of the resulting sound. Note the highly dynamic nature of the very beginning of the signal, a phenomenon that confers a great degree of realism to the result. This is very difficult to obtain with other synthesis methods such as subtractive synthesis. The center graph (Frontispiece b) plots the excitation pressure at the bottom region of the trachea section of the model, and the bottom graph (Frontispiece c) shows the behavior of the vocal folds in terms of the width of the glottis during vibration. After a rapid settling period, the glottis vibrates steadily in response to the regular pressure.

Figure 3 shows the cochleogram of a sequence of four different vowels produced by Robovox. Note the dynamic behavior of the formants, represented by the lines running through the shaded areas.

Additive Re-Synthesis Example

Each sound in Diphone is referred to as a segment. The program produces a set of analysis data for each segment and stores it in a dictionary of segments. Diphone provides an intuitive user interface for operations, in which one can drag segments from a dictionary and drop them into a working area, referred to as the sequence window. At the top of the sequence window the computer displays the segments for concatenation and at the bottom it provides a menu of parameters for monitoring the concatenation.

The program concatenates the sounds by applying an algorithm that interpolates the analysis data of neighboring segments. It is important to stress that Diphone does not manipulate the sounds directly, only the analysis data (Fig. 4).

Composers working with recorded sound normally concatenate the sounds using splicing and cross-fading techniques. In most cases, however, the results do not sound satisfactory because the inner contents of the sounds involved do not always match. Diphone is a good tool for forging such concatenations.

The following example shows the creation of an artificial utterance combining four segments extracted from words spoken in Irish and Croatian. Four utterances were used, two from each language (Figs 5 and 6).

The segments were combined in sequence and the adjacent areas were interpolated accordingly. The result is shown in Fig. 7. A number of trials were made in order to find optimal interpolations. The criteria for assessing the results were subjective, the main requirement being that the utterance should sound as realistic as possible.

Prosody Extraction Example

As an example of prosody extraction, consider the utterance "nineteen fifty-five" as spoken by an adult female British speaker (Fig. 8).

The extracted pitch envelope is represented in Fig. 9a, and the energy contour plotted as an intensity graph in dB is shown in Fig. 9b. The pitch of this utterance ranges from approximately 240 Hz to 420 Hz: it rises almost exponentially from approximately 280 Hz up to 420 Hz at the syllable /fi/ of the word "five." Then, it decays as low as 240 Hz and rises again toward 420 Hz at the end of this word. Note that there is a gap in the pitch line just before the syllable /ty/. This is due to articulatory phenomena. As the consonant /t/ of the syllable /ty/ is a voiceless stop, the vocal folds stop vibrating for the production of this plosive; hence there is no pitch. In this case, a rise in subglottal pressure, combined with a sudden stiffness of the vocal folds just after the stop, cause the sudden rise in pitch and amplitude that culminated in the stressed fricative syllable /fi/.

The result of the synthesis stage is shown in Fig. 10a. By giving a brief glance at this representation of the signal, one can immediately infer that this synthesized utterance lacks the dynamics of a spoken sound. If we listened to it we would certainly be able to follow the pitch variation shown in Fig. 9a, but the utterance sounds rather unnatural, despite its convincing vocal timbre. Hence the importance of the last stage of the PROSE algorithm: the modulation. The outcome of the modulation is represented in Fig. 10b. Compare the dynamics of this sound with the (lack of) dynamics of the one represented in Fig. 10a.

Prosody Manipulation Example

Whereas PROSE extracts prosody from an utterance, PSOLA is a powerful tool for modifying its prosody. The example below illustrates an example in which the utterance's pitch and rhythm have been manipulated. The utterance in question has three syllables. Its original spectrum is shown in Fig. 11a, and its pitch analysis is shown in Fig. 11b.

The pitch sequence of the utterance was reversed, and the duration of the

first and second syllables were slightly changed; the first was shortened and the second was enlarged. Figure 12b shows the pitch analysis of the modified utterance. Also visible in this figure are the changes in the duration of the first two syllables. Figure 12a shows the spectrogram of the new utterance. Note that the spectrum content (frequency domain) remains identical to the spectrum content of the original sound. This is the power of the PSOLA technique: one can manipulate the prosody of the utterance without damaging the timbre.

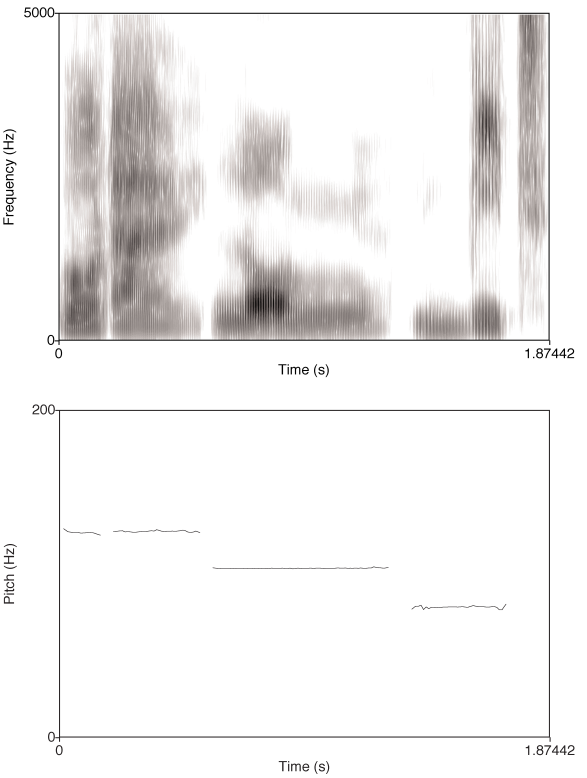


Fig. 12. (a) The spectrogram of the new utterance with pitch sequence reversed. (b) Pitch analysis of the new utterance with reversed pitch sequence. (© Eduardo Reck Miranda)

Concluding Remarks

My compositional activities are deeply rooted in the belief that experimental but highly methodical approaches to musical composition are powerful methods for hands-on research into sound systems and cognition. The composition of this piece helped me to gain an unprecedented practical understanding of the human voice, its computer simulation and cognition.

In many ways, this piece is an attempt to extend the lexicon of vocal music to a new dimension, combining surreal linguistic systems and synthesized and computer-manipulated voice.

The more sophisticated the technique used to synthesize and/or process voice, the more realistic the results obtained, but this sophistication often comes at a price: It requires more know-how and indeed patience to operate. Physical modeling synthesis of the human voice is very promising, but it is very difficult to control. Also, it is computationally very intensive, making it almost impossible for real-time synthesis even on the most powerful computers. The advantage of working with a physical model is that its own "physiology" often guarantees realistic results. Even though one may program the model to behave in ways that would be humanly impossible, the results often sound very much like the human voice.

I have deliberately avoided working with languages that I can speak fluently in order to prevent biased results. However, it has not been possible to completely avoid this bias. As we grow up our brains develop themselves to deal efficiently with the languages that surround us [14]. The brain develops listening preferences and speech motor abilities that are very difficult to change in adulthood. Thus, the older we get, the more difficult it is to learn a new language without the accent of our mother tongue. Therefore my auditory system has most probably acted as a filter in my selection of the results from the various synthesis practices to be included in the piece. This is a plausible explanation for the fact that only approximately 20% of my attempts at creating artificial utterances were satisfactory to my ears. Nevertheless, the results obtained with Diphone still sounded more realistic than what I would have produced with these utterances by splicing the individual segments sequentially with signal cross-fading.

As for prosody extraction and manipulation, PROSE and PSOLA proved to be very powerful. They opened an important window for experimenting with the musicality of the various languages used in the piece.

Acknowledgments

This composition was achieved thanks to a John Simon Guggenheim Foundation Fellowship awarded to the author in 2000. Part of the composition was realized at Tempo Reale in Florence and part at CCMIX in Paris between 2000 and 2002 [15].

References and Notes

1. P. Ladegofed and I. Maddieson, *The Sounds of the World's Languages* (Oxford, U.K.: Blackwell, 1996).

2. IPA, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet* (Cambridge, U.K.: Cambridge Univ. Press, 1999).

3. N.H. Fletcher and T.D. Rossing, *The Physics of Musical Instruments* (Berlin: Springer-Verlag, 1998).

4. D.H. Klatt, "Software for a Cascade/Parallel Formant Synthesizer," *Journal of Acoustic Society of America* **67**, No. 3, 971–995 (1980).

5. E.R. Miranda, *Computer Sound Design: Synthesis Techniques and Programming* (Oxford, U.K.: Elsevier/Focal Press, 2002).

6. J. Sundberg, *The Science of the Singing Voice* (DeKalb, IL: Northern Illinois Univ. Press, 1987).

7. P. Boersma, "Synthesis of Speech Sounds from a Multi-Mass Model of the Lungs, Vocal Tract, and Glottis," *Proceedings of the Institute of Phonetic Sciences Amsterdam,* No. 15 (1991) pp. 79–108.

8. X. Serra, "Musical Sound Modeling with Sinusoids Plus Noise," C. Roads et al., eds., *Musical Signal Processing* (Lisse, The Netherlands: Swets & Zeitlinger, 1997) pp. 91–122.

9. C. Dodge and T.A. Jerse, *Computer Music: Synthesis, Composition, and Performance* (New York: Schirmer Books, 1997).

10. X. Rodet and A. Lefevre, "Macintosh Graphical Interface and Improvements to Generalized Diphone Control and Synthesis," *Proceedings of the International Computer Music Conference* (ICMC'96), Clear Water Bay, Hong Kong (San Francisco: ICMA, 1996).

11. E.R. Miranda, "Synthesising Prosody with Variable Resolution," *Proceedings of the 110th Convention of the Audio Engineering Society* (AES), Amsterdam, The Netherlands (New York: Audio Engineering Society, 2001).

12. C. Roads, *The Computer Music Tutorial* (Cambridge, MA: MIT Press, 1996).

13. H. Valbret, E. Moulines and J.P. Tubach, "Voice Transformation Using PSOLA Technique," *Speech Communication* **11**, Nos. 2–3, 175–187 (1992).

14. D. Drubach, *The Brain Explained* (Upper Saddle River, NJ: Prentice-Hall, 2000).

15. *Sacra Conversazione* exists in two forms: as a short opera with human singers (approximately 25 minutes long) and as an entirely electroacoustic piece (approximately 15 minutes long). The recording of the premiere of the former version is not publicly available at the time of writing. Excerpts of the electroacoustic-only version and other sound exam-

ples, including those mentioned in this paper, are available on request. Alternatively, please refer to <cmr.soc.plymouth.ac.uk/lmj15_examples/>.

Discography of Works for Voice

Groupe Vocal de France (performers), *Messiaen & Xenakis,* Arion CD ARN68084 (1989).

Linda Hirst (mezzo-soprano), *Songs Cathy Sang,* Virgin Classics CD VC790704-2 (1988).

Eduardo Reck Miranda, *Mother Tongue,* Sargasso CD SCD28051 (2004).

Schola Heidelberg and Ensemble Aisthesis (performers), *Nuits—weiss wie Lilien,* BIS Records CD BIS-CD-1090 (2001).

Sinopoli—Staatskapelle Dresden (performers), *Schoenberg: Pierrot Lunaire—Erwatung,* Teldec CD 3984-22901-2 (1999).

Various and anonymous, *Voix de L'orient Sovietique,* Inedit—Maison Cultures du Monde CD W260008 (1989).

Various composers, *Computer Music Currents* ***1,*** Wergo CD WER2021-50 (1989).

Various composers, *Computer Music Currents* ***4,*** Wergo CD WER2024-50 (1989).

Various composers, *Computer Music Currents* ***7,*** Wergo CD WER2027-50 (1990).

Trevor Wishart, *Tongues of Fire,* Trevor Wishart CD OTP001 (1994).

Manuscript received 1 January 2005.

Eduardo Reck Miranda holds a Ph.D. in music from the University of Edinburgh. He served as a research scientist for Sony for a number of years, conducting research on speech technology and evolution of language. He currently is head of the Computer Music Research department at the University of Plymouth, U.K. He has authored a number of patented works, papers and books in the field of computer music and has made contributions in the fields of speech, evolutionary music and cognitive neural modeling. His latest CD, Mother Tongue, *was released by Sargasso in London in 2004. He is the regional editor for Latin America of* Organised Sound *and a member of the editorial boards of* Leonardo Music Journal *and* Contemporary Music Review.