Computer-Aided Song Design: Prosody as Scaffolding

Eduardo Reck Miranda
SONY Computer Science Laboratory Paris
6 Rue Amyot – 75005 Paris – France
miranda@csl.sony.fr - http://www.csl.sony.fr

Abstract

Composers often face the task of composing melodies for given texts such as religious hymns, poetry or the libretto of an opera. A plausible point of departure for writing melodies for a text is to study the prosody of the text as spoken either naturally or dramatically. In this paper we introduce PROSE: a system for aiding such study. The system extracts the prosody of a spoken signal and (re)synthesises it at various resolutions. The main advantages of using PROSE over simply listening to the spoken signals are that composers can focus on prosodic auditory information detached from the meaning of the text and can assess this information at various resolutions. Also, the analysis data can be plotted for visual assessment and/or mapped onto musical parameters.

1. Introduction

Composers often face the task of composing melodies for given texts such as religious hymns, poetry or the libretto of an opera. Different composers have approached this problem in a variety of ways, ranging from sheer inspiration and intuition to formal approaches. Formal approaches to song writing in Western European music are as old as music theory itself. One of the finest examples of early formal approaches to song writing appeared in the eleventh century, when Guido d'Arezzo proposed a lookup chart for assigning pitch to the syllables of religious hymns. He also invented the musical stave for systematic notation of music and established the medieval scales known as the Church modes (Sadie 1981). Although computer music systems are normally orientated towards formal approaches to composition (Miranda 2001), in this paper we propose the use of the computer as a powerful aid to inspiration, rather than as a processor of formalisms for generative song writing.

Writers and poets often assess their work by reading them aloud: in these cases a text is finally right when its intonation *sounds right*. What does "sounds right" mean here? What makes the placement of a word "x" in a sentence sounds better than a word "y" of identical lexical meaning? There are no simple answers to these questions. Indeed, these are hot topics in current Linguistic research and surveys of the intonation systems of various languages using a common methodology have already begun to appear (Hirst and DiCristo, 1998). Still, very little is known about the overall intonation properties of

different languages. What we do know, however, is that prosody seems to hold an important key to unlock these questions.

Composers, on the other hand, do not usually write the texts for their songs, but work with given texts. Some texts are perhaps easier than others to work with because their intonation properties were better explored by the writer. A plausible point of departure for writing melodies for a text is therefore to study the prosody of the text as spoken either naturally or dramatically.

What do we mean by "prosody"? How can the computer aid composers to study the prosody of a text? The literature on phonetics often defines prosody primarily in terms of the pitch variations of an utterance (Clark and Yallop, 1990). On a few occasions it is rhythm that has been considered to be the predominant attribute of prosody (Rasmus *et al.*, 1999). For the purpose of this paper, however, we purport the notion that *intonation* of utterances and *voice timbre* are two important determinants of prosody (Hepper *et al.*, 1993) (Goodman and Nusbaum, 1994) (Lecanuet and Werker, 1995) (Jusczyk, 1997). As for the second question, the computer can aid the composer by providing means for listening and visualising the prosody of a speech signal, detached from the semantic meaning of the words. Also, once the prosodic-specific information has been extracted from the signal, the computer can map this information onto specific musical parameters for fast-prototyping musical ideas.

In order to build such a system, one must establish the acoustic correlates of intonation and timbre in order extract this information from the actual speech signal. In this case, the acoustic correlates of intonation are the *pitch* and *amplitude* unfolding of the signal, whereas the acoustic correlates of voice timbre are its first five *formants*. Once these correlates have been extracted, the system should then be capable of: a) (re)synthesising the prosody using these values, b) plotting the analysis values for visual assessment and c) mapping these values onto musical parameters (e.g., pitch contour could be used to generate note sequences). An additional desired feature of such a system is the ability to zoom in and out the analysis resolution; the advantages of being able to do this will be discussed later.

In the remaining sections of this paper we introduce PROSE (short for PROSody Extractor): a system for aiding composers to study prosody. Basically, PROSE extracts the prosody of a spoken signal and synthesises it at various resolutions. The extracted data can be plotted for visual assessment and mapped onto musical parameters. This paper introduces the analysis and synthesis algorithms that allows for synthesising the prosody and plotting the analysis values for visual assessment. The third capability, that is, the mapping of the analysis data onto musical parameters, is the topic for a forthcoming paper.

2. The PROSE system

The overall algorithm of PROSE is summarised in Figure 1. The system starts by extracting the pitch envelope and the energy contour of a speech stream. Then, the extracted pitch envelope is fed into a source-filter speech synthesiser that produces an open vowel sound with F_0 variations; the input pitch envelope drives these F_0 variations. Following that, 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.

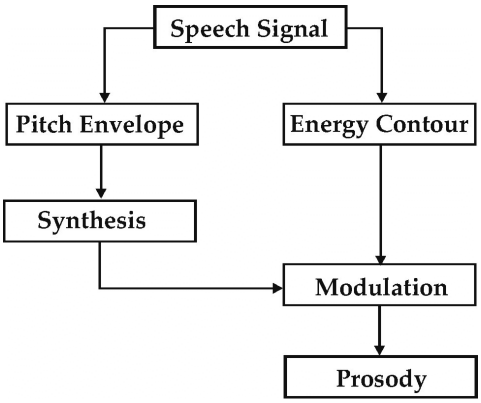

Figure 1: Block diagram of the algorithm for PROSE.

The pitch extraction module employs an improved autocorrelation-based technique proposed by Boersma (1993). Basically, 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 *n* 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. Autocorrelation is generally defined as follows:

$$r_x(\tau) = \sum_{i=0}^{I} x(i)x(i+\tau) \qquad (1)$$

where I is the length of the sound stream in terms of samples, $r_x(\tau)$ is the autocorrelation as a function of the lag τ, $x(i)$ is the input signal at sample i, and $x(i+\tau)$ is the signal delayed by τ, such that $0 < \tau \leq I$. The magnitude $r_x(\tau)$ is given by the degree to which the value of $x(i)$ is identical to itself delayed by τ. Therefore the output of the autocorrelation gives the magnitude for different lag values. In practice, the function

$r_x(\tau)$ has a global maximum for $\tau = 0$. If there are global maximums beyond 0, then the signal is periodic in the sense that there will be a time lag T_0 so that all these maximums will be located at the lags nT_0, for every integer *n*, with $r_x(nT_0) = r_x(0)$. The frequency of this signal is calculated as $F_0 = 1/T_0$.

Note that the equation (1) assumes that the signal *x*(*i*) is stationary, but speech is a highly non-stationary signal. In this case, a short-term autocorrelation analysis is forged by windowing the signal. This gives estimates F_0 at different times. The pitch envelope of the signal *x*(*i*) is obtained by placing a sequence of $F_0(t)$ estimation for various windows *t* in an array *P*[*t*]. The algorithm uses a Hanning window (Ramirez, 1985) whose length is determined by the lower frequency value candidate that we would expect to find in the signal.

The synthesis task is performed by a source-filter synthesis architecture. A source-filter synthesiser is based on the idea that the production of vocal sounds can be simulated by treating it as the generation of some type of raw source sound which subsequently passes through a filter arrangement (Miranda, 1998). In humans, the raw sound source would correspond to the outcome from the vibrations created by the vocal folds and the filter arrangement to the vocal tract (Klatt, 1980).

The implementation of the filter arrangement is based upon measurements of the human vocal tract. In general, the vocal tract is considered as a tube (with a side-branch for simulating the nose), sub-divided into a number of cross-sections whose individual resonance is simulated by a filter. The outcome Φ(*f*) of the source-filter synthesiser can be characterised in the frequency domain as follows:

$$\Phi(f) = S(f)\Delta(f) \tag{2}$$

where S(*f*) is a source signal and Δ(*f*) is a linear transfer function defined by the filter arrangement. Given an input signal $x(n\vartheta)$, such as a low-pass filtered pulse train, and a constant ϑ equal to the inverse of the sampling rate, the filter arrangement is composed of a combination of digital resonators as follows:

$$\delta(n\vartheta) = Ax(n\vartheta) + B\delta(n\vartheta - \vartheta) + C\delta(n\vartheta - 2\vartheta) \tag{3}$$

where $\delta(n\vartheta - \vartheta)$ and $\delta(n\vartheta - 2\vartheta)$ are the two previous samples of the output $\delta(n\vartheta)$. The values of *A*, *B* and *C* are specified according to the resonance centre frequency F_c and bandwidth *W* values of the desired formant, as follows:

$$C = -e^{(-2\pi W\vartheta)}$$
$$B = 2e^{(-\pi W\vartheta)}\cos(2\pi F_c\vartheta) \quad (4)$$
$$A = 1 - B - C$$

The synthesiser uses five such digital resonators in parallel, each of which produces one formant; there is no provision for nasal speech in this model. Since we only want to produce a single open vowel throughout the utterance, the values of F_c and W can be settled as shown in Table 1. The speech timbre therefore results from the addition of the signals produced by each resonator δ_n scaled by an attenuation coefficient γ_n that defines the amplitude of each formant in the overall spectrum:

$$\Phi(t) = \gamma_1\delta_1(t) + \gamma_2\delta_2(t) + ... + \gamma_5\delta_5(t). \quad (5)$$

	F_c	W	γ
Resonator 1	622.25 Hz	60 Hz	0 dB
Resonator 2	1568 Hz	90 Hz	-7 dB
Resonator 3	2489 Hz	120 Hz	-9 dB
Resonator 4	3400 Hz	250 Hz	-12 dB
Resonator 5	4500 Hz	350 Hz	-22 dB

Table 1: Centre frequency, bandwidth and attenuation values for the resonators.

The energy contour is obtained by convolving the squared values of the samples with a smooth bell-shaped curve with very low peak-side lobe (e.g., -92 dB or less). Convolution can be defined as follows:

$$\varepsilon(k) = \sum_{n=0}^{N-1} x(n)^2\,\upsilon(k-n) \quad (6)$$

where $x(n)^2$ represents a squared sample n of the input signal x, N is the total amount of samples in this signal and k ranges over the length of the window υ. The length of the window is set to one and a half times the period of the average fundamental frequency. (The average fundamental frequency is obtained by averaging the values of pitch envelope $P[t]$ obtained earlier). In order to avoid oversampling the contour envelope, only the middle sample value for each window is convolved. These values are normalised after convolution for further processing.

Finally, the modulation stage employs the amplitude modulation technique whereby the synthesised utterance is multiplied by a normalised version of the energy contour:

$$P(t) = \Phi(t)\varepsilon(t) \tag{7}$$

where P(t) is the prosody result at sample t; $\Phi(t)$ is the synthesised carrier signal and $\varepsilon(t)$ is the modulating energy contour. In this case, no side-bands are added to the spectrum of the carrier because the modulating signal is infrasonic.

3. Example of prosody extraction and synthesis

Before we go on to present the adaptable features of PROSE, let us examine an example of prosody extraction and synthesis. Consider the utterance "nineteen fifty-five" as spoken by an adult female British speaker (Figure 2).

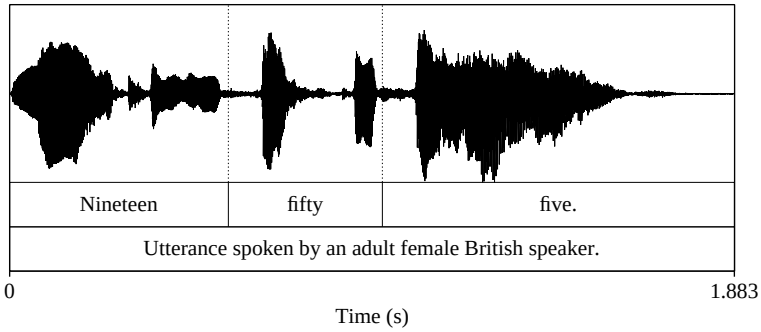

Figure 2: The time domain representation of a speech stream.

The extracted pitch envelope is represented in Figure 3 and the energy contour plotted as an intensity graph in dB is shown in Figure 4. The pitch in this utterance ranges from approximately 240 Hz to 420 Hz: its rises 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 towards 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 (Ladefoged and Maddieson, 1996), the vocal folds stop vibrating for the production of this plosive; hence there is no pitch. In this case, a rise in sub-glottal pressure, combined with a sudden stiffness of the vocal folds just after the stop, causes the sudden rise in pitch and amplitude that culminates in the stressed fricative syllable /fi/. The observant reader will probably ask at this point why PROSE did not capture a similar phenomenon that occurred before the syllable /teen/ of the word "nineteen"? This is a property of the analysis technique that will be clarified in the following section when we discuss the adaptable features of our model. At the moment it suffices to say that the transition from the nasal ending of the segment /nine/ to the stop

[t] is different from the transition from the fricative ending of the syllable /fif/. The vocal folds stop vibrating for a longer time lapse in the latter case.

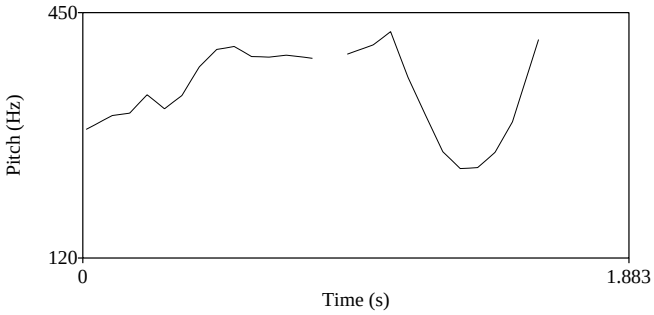

Figure 3: The extracted pitch envelope of the utterance shown in Figure 1.

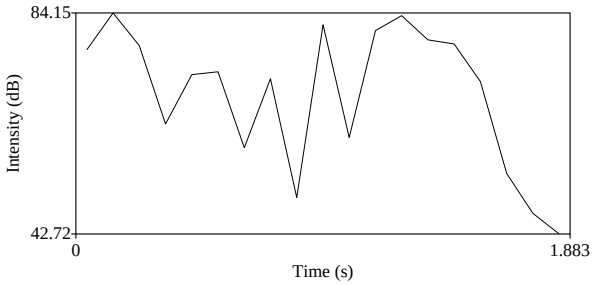

Figure 4: The energy contour of the utterance shown in Figure 1.

The intensity graph complements our reading of the behaviour of the vocal system, but from another angle. Here we can infer the movements of the vocal tract, most notably those that obstruct the sound flow coming from the glottis: such obstructions are marked by amplitude decays. Such decays can indicate syllable boundaries, but this is not always the case. Note that the word “five”, for example, has two syllables, but the graph in Figure 4 does not indicate this. From a musical point of view, however, this is by no means a fault. On the contrary, people do not hear two separate syllables here, but a single segment, phonetically represented as [faIv].

The result of the synthesis process is shown in Figure 6. By giving a brief glance at this representation of the signal, one can immediately infer that this synthesised utterance lacks the dynamics of a spoken sound. If we listen to it we would certainly be able to follow the pitch variation shown in Figure 4 but the utterance sounds rather unnatural, despite its highly convincing vocal timbre. Hence the importance of the last stage of the PROSE algorithm: the modulation.

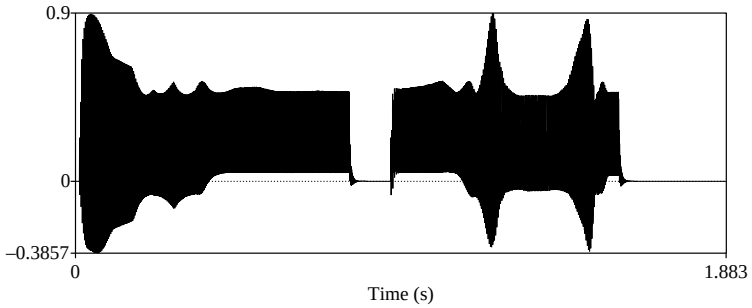

Figure 5: The synthesis result prior to modulation.

The outcome of the modulation stage is represented in Figure 6. Compare the dynamics of this sound with the (lack of) dynamics of the one represented in Figure 5. If we listen to it we will not only be able to follow the energy contour of Figure 3, but we will notice that of the whole prosody sounds much more convincing to our ears: the pitch variation is much clearer. Also note that PROSE has captured the distribution of stress in the utterance well; this is important for studying the rhythm of a text.

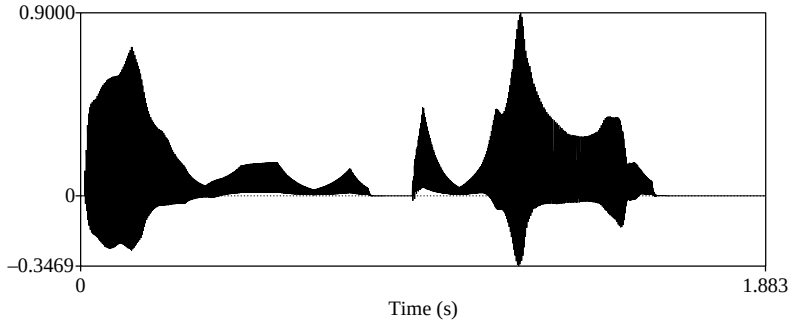

Figure 6: The synthesis outcome from PROSE.

4. The zooming feature

As discussed earlier, it is useful to render our prosody analysis/synthesis system adjustable to different resolutions. This is achieved by adjusting the resolution of the procedures for extracting the pitch envelope and the energy contour. These resolutions are given by the time-step of the windowing processes for the application of the equations (1) and (6), respectively.

The length of the window for the short-term autocorrelation analysis for the examples in this paper was set to the inverse of the minimum frequency that the algorithm would try to match in the sound; e.g.120 Hz. It is assumed by default that adult women do not speak lower than 120 Hz and that adult men do not speak lower than 50 Hz. This value can, of course, be changed to suit specific cases such as the voice of a teenager or a child. The time-step for the pitch-extraction example shown in Figure 4 is

equal to 0.06; this is considered as a middle resolution analysis. A high-resolution analysis can be achieved by setting this value to 0.02 and a much lower resolution can be achieved by setting this value as high as 0.1.

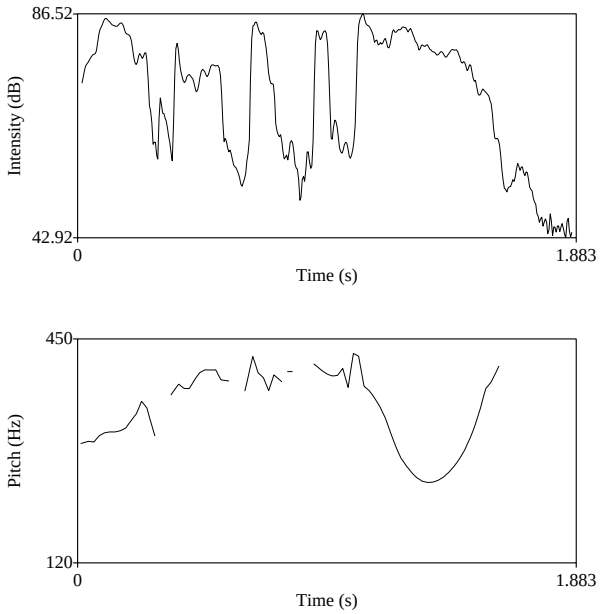

Figure 7: An example of high-resolution analysis.
The top graph shows the energy contour and the bottom graph the pitch envelope.

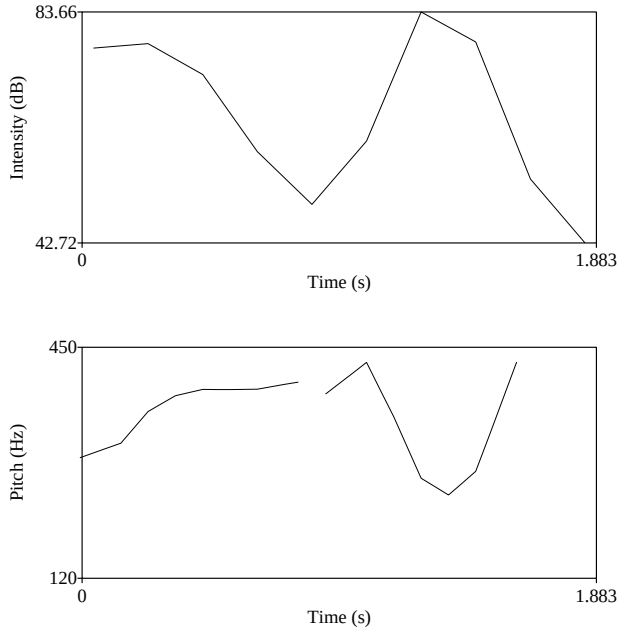

Figure 8: An example of low-resolution analysis:
energy contour at the top and pitch envelope at the bottom.

As for the resolution of the energy contour, the time-step for the example shown in Figure 4 is equal to 0.1. This yields to a middle resolution analysis. A high-resolution contour can be obtained by setting the time-step equal to the inverse of the average fundamental frequency of the utterance. Conversely, a value equal to 0.2 will yield a rather low-resolution contour.

The outcome of the higher and lower settings mentioned above can be seen in Figures 7, 8 and 9. Figure 7 illustrates the case of high-resolution energy contour (top) and pitch envelope (bottom) analyses, whereas Figure 8 illustrates the low-resolution cases. The synthesis result for both cases is shown in Figure 9: high-resolution at the top and low-resolution at the bottom of the figure.

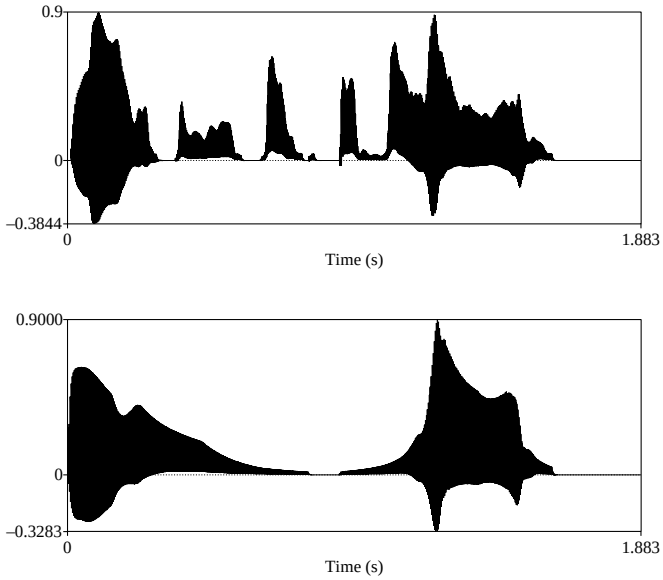

Figure 9: The outcome of the PROSE system:
high-level prosody extraction at the top and low-level at the bottom.

5. Conclusion

In this paper we presented PROSE, a system for aiding composers to study the prosody of spoken texts based upon information extracted from the acoustic signal. The overall algorithm of PROSE starts by extracting the pitch envelope and the energy contour of the given signal. The extracted pitch envelope is fed into a source-filter speech synthesiser that produces a vowel sound with variations in fundamental frequency. Then, the result is amplitude modulated by the extracted energy contour. The extracted information can be plotted for visual assessment and/or mapped onto musical parameters.

The main advantages of PROSE over simply listening to the spoken signal are: a) composers can focus on the prosodic auditory information, detached from the meaning of the utterances and b) composers can assess prosody at various resolutions. PROSE also

proved to be useful for electroacoustic music, where the synthesised prosody can be used directly as source materials for compositions.

At present, the synthesised prosody has an open vowel-like timbre. This certainly gives vocal quality to the sound, but the timbre of the outcome should ideally follow the timbral unfolding of the original speech. We are currently improving PROSE by adding a formant extraction module to the algorithm shown in Figure 1. The extracted formant information will be used at the synthesis stage in order to mimic the formant evolution of the original signal. The main burden of designing such a formant analyser concerns the resolution of the zooming requirement discussed in the introduction. In contrast to pitch envelope and energy contour, the degree of competence in formant listening/production cannot be simulated by tweaking the time resolution of the analysis windowing procedure. Formant resolution here involves the ability to distinguish between timbres in a vowel space and this involves categorisation mechanisms that are not straightforward for simulating with standard signal processing techniques.

The author has successfully used PROSE to compose sections of the eletroacoustic and the vocal parts of the piece *Sacre Conversazione*, for vocal quartet, electroacoustic support and live electronics, commissioned by the John Simon Guggenheim Memorial Foundation.

References

Boersma, P., 1993. "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound", *University of Amsterdam IFA Proceedings*, No. 17, pp. 97-110.

Clark, J. and Yallop, C., 1990. *An Introduction to Phonetics and Phonology*. Oxford (UK): Blackwell.

Goodman, J. C. and Nusbaum, H. C. (Eds.), 1994. *The Development of Speech Perception*. Cambridge (MA): The MIT Press.

Hepper, P. G., Scott, D., and Shahidullah, S., 1993. "Newborn and fetal response to maternal voice", *Journal of Reproductive and Infant Psychology*, No. 11, pp. 147-153.

Hirst, D. and Di Cristo, A. (Eds.), 1998. *Intonation Systems: A Survey of Twenty Languages*. Cambridge (UK): Cambridge University Press.

Jusczyk, P., 1997. *The Discovery of Spoken Language*. Cambridge (MA): The MIT Press.

Klatt, D. H., 1980. "Software for a cascade/parallel formant synthesizer", *Journal of the Acoustical Society of America*, Vol. 67, No. 3, pp. 971-995.

Ladefoged, P. and Maddieson, I., 1996. *The Sounds of the World Languages.* Oxford (UK): Blackwell Publishers.

Lecanuet, J-P., and Granier-Deferre, C., 1993. "Speech stimuli in the fetal environment", B. de Boysson-Bardies *et al.* (Eds.), *Developmental Neurocognition: Speech and*

Facing Processing in the First Year of Life. Dordrecht: Kluwer Academic Publishers

Locke, J. L., 1993. *The Child's Path to Spoken Language*. Cambridge: Harvard University Press.

Miranda, E. R., 2001. *Composing Music with Computers*. Oxford (UK): Focal Press.

Miranda, E. R., 1998. *Computer Sound Synthesis for the Electronic Musician*. Oxford (UK): Focal Press.

Nazzi, T., Floccia, C., and Bertoncini, J., 1998."Discrimination of pitch contours by neonates", *Infant Behaviour,* No. 21, pp. 543-554.

Ramirez, R. W., 1985. *The FFT - Fundamentals and Concepts*. Englewood Cliffs (NJ): Prentice-Hall.

Rasmus, F., Nespor, M. and Mehler, J., 1999. "Correlates of linguistic rhythm in the speech signal", *Cognition,* No. 73, pp. 265-292.

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

van Santen, J. P.H., Sproat, R.W., Olive, J.P., Hirschberg, J. (Eds.), 1997. *Progress in Speech Synthesis*. NewYork: Springer Verlag.

Sadie, Stanley, ed. *The New Grove Dictionary of Music and Musicians.* New York (NY): Macmillan, 1980-1981.