The role of speech synthesis in *Requiem per una veu perduda*

EDUARDO RECK MIRANDA

SONY Computer Science Laboratory, 6 rue Amyot, 75005 Paris, France
E-mail: miranda@csl.sony.fr

***Requiem per una veu perduda* (*Requiem for a lost voice* in Catalan) is a piece for mezzo-soprano with on-stage effects processing and recorded electroacoustic material. The electroacoustic material consists mainly of synthesised vocal sounds and the pitches for the mezzo-soprano part were defined based upon the functioning of the speech synthesis methods used. This paper outlines the basics of one of the speech synthesis methods used to compose the piece and introduces the scheme for the definition of the pitch material for the singing. The relationship between the speech synthesis methods and the pitch material establishes the structural foundations of the piece.**

1. INTRODUCTION

Requiem per una veu perduda (*Requiem for a lost voice* in Catalan) is a piece for mezzo-soprano with effects (pitch shift, reverb, etc.) and recorded electroacoustic material (either on DAT or CD). The latter is fixed, i.e. it does not change during performance (sic), but the soprano has some flexibility for synchronising her singing with this material. The piece was commissioned by Phonos/Pompeu Fabra University's Audiovisual Institute (IUA), Barcelona, and it was completed in July 1997. So far, it has been performed live in Brazil, the Czech Republic, Cuba, Scotland and Spain, and broadcast in Argentina by *Radio de la Cuidad de Buenos Aires*. The Requiem has four movements of approximately four minutes each: *Kyrie*, *Gloria*, *Sanctus-Benedictus* and *Agnus Dei* (the last two movements are available on the accompanying journal CD).

The electroacoustic material was produced using three types of synthesis techniques: (i) physical modelling, (ii) spectral modelling and (iii) granular synthesis (Miranda 1998a). For granular synthesis, the author used Chaosynth, a software of his own design (Miranda 1995a); the *granular* sounds can be heard at various sections in the piece.

Spectral modelling was used to modify either or both the form and the content of the spectrum of various sampled sounds, mostly from Catalan poetry readings. The systems used here were the SMS package (Serra 1997) and CDP's phase vocoder (Fishman 1997). However, it is the physical modelling technique that is the main focus of this paper, as this was used to synthesise the Latin lyrics of the *Gloria* and the speech-like sounds of the entire piece.

The pitch material for the mezzo-soprano part was defined based upon the functioning of the speech synthesis methods used. This paper outlines the basics of one of the speech synthesis methods used to compose the piece and introduces the scheme for the definition of the pitch material for the singing. The relationship between the speech synthesis methods and the pitch material establishes the structural foundations of the piece.

1.1. Inspiration

The relationship between the sonority of the human voice and its potential to express emotions and ideas, not necessarily within the context of a specific language, has always fascinated the author. People who have visited a foreign culture, with no knowledge of its language, often acknowledge that they could understand a limited repertoire of paralinguistic noises independent of the speaking language. This phenomenon also seems to be closely related to our impulsion towards music-making. In this piece the author wanted to experiment with this relationship.

The notion that our linguistic capacity is closely related to our ability to both make and appreciate music was prominent in Enlightenment thinking in the eighteenth century (Thomas 1995). Reflections on how primordial utterances, cries and vocalisations would have evolved into language naturally brought musical considerations within the scope of the writings of philosophers such as Condillac and Rousseau.

Condillac in his *Essay on the Origins of Human Knowledge* depicted the earliest spoken language as being composed of action-orientated vocal inflections such as warnings, cries for help, shouts of joy, etc. But most interestingly, Condillac proposed that these inflexions were accompanied by variations in pitch and timbre. In short, he suggested that the early hominids did not prioritise the invention of different 'words', but tended to produce the same form of utterance at different tones in order to express different things; presumably by varying pitch, loudness and duration. Condillac thus suggested that primordial

Organised Sound **3**(3): 235–40 © 1998 Cambridge University Press. Printed in the United Kingdom.

https://doi.org/10.1017/S1355771898003070 Published online by Cambridge University Press

languages did not have consonants but vowel-like intonations. The prosody of earlier languages must have sounded like a kind of primitive song (Arbo 1998).

Rousseau also purported the idea that language is derived from natural sounds produced by our vocal organs. For Rousseau, however, song and speech have a common ground: *passion*. In the beginning, vocal utterances primarily expressed feelings (e.g. 'I am sad'), whilst gestures were normally preferred to express rational thoughts (e.g. 'Go hunting, I am hungry!'). Rousseau agrees with Condillac that primeval spoken languages must have sounded like melodies of vowel-like utterances, but Rousseau has an interesting story for the emergence of consonants: as hominids' dealings with one another grew in complexity, spoken language needed to become less passionate and more precise. In his *Essay on the Origin of Language*, Rousseau argues that language was motivated by the increasing necessity for social bonding. Within this bonding process, the amount of tone variations decreased, giving rise to the appearance of consonants. New articulations needed to be formed, and consequently grammatical rules for making sequences of utterances soon emerged. For Rousseau, modern languages (such as his own mother tongue French) no longer spoke to the heart alone, but also to reason. As language followed the path of logical argumentation, those melodic aspects of the primordial utterances evolved into music instead. Music thus developed from the sounds of passionate speech.

Since Condillac and Rousseau, the relationship between the origins of language and music has hardly been systematically addressed again. Their writings seem to have been overshadowed by the romanticism that prevailed in Europe in the nineteenth century, most notably in music; influential philosophers who were interested in music often associated its origins with the mystical, the ineffable and the hidden. Indeed, in the 1880s, the prestigious Linguistic Society of Paris sought to impose a ban on the theme of the origins due to a wave of wild unsupported writings that appeared at the time.

To compose *Requiem per una veu perduda*, the author also sought inspiration from the beauty of Catalan poetry, the joy of spontaneous singing, the power of worship and the magic of artificial voice synthesis. It is a requiem for our own voice; our errant, quixotic voice that we often leave aside in favour of the rational grammar of a language.

2. SYNTHESISING HUMAN VOCAL SOUNDS USING SUBTRACTIVE SYNTHESIS

The spectral contour of vocal-like sounds has the appearance of a pattern of 'hills and valleys' technically called formants (figure 1). The centre frequencies of the first three formants (f_1,f_2 and f_3) are crucial for the identity of the sound. Among the various synthesis techniques capable of synthesising formants, both the *subtractive* and the *waveguides* synthesis techniques were used. This paper focuses on the former, as it is the technique used by ARTIST, the computer program that synthesised most of the vocal-like sounds of the piece (Miranda 1995b, 1998b).

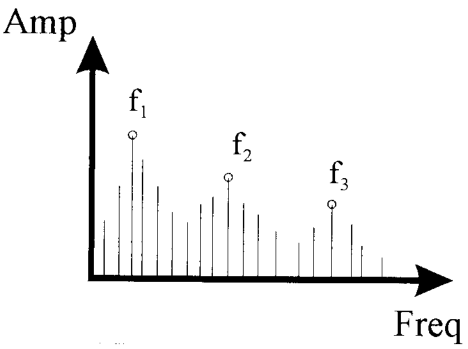

Figure 1. The spectral contour of vocal-like sounds has the appearance of a pattern of 'hills and valleys' known as formants.

In subtractive synthesis each formant is associated with the response of a band-pass filter (BPF); in this case, one needs a parallel composition of BPFs set to different responses. The signal to be filtered is simultaneously applied to all filters and the frequency responses of the filters are added together. This synthesis approach considers that the behaviour of the human vocal tract is determined by two main components: *source* and *resonator* (or *filter*), where the former produces a raw signal that is shaped by the latter (figure 2).

When singing or speaking, an airstream is forced upwards through the trachea from the lungs. At its upper end, the trachea enters the larynx, which in turn opens into the pharynx. At the base of the larynx, the vocal folds are folded inwards from each side, leaving a variable tension and a slit-like separation, both controlled by muscles in the larynx. In normal breathing, the folds are held apart to permit the free flow of air. In singing or speaking, the folds are brought close together and tensed. The forcing of the airstream through the vocal folds in this state sets them into vibration. As a result, the airflow is modulated at the vibration frequency of the vocal folds. Despite the fact that the motion of the vocal folds is not a simple but a nonuniform vibration, the pitch of the sound produced is determined by this motion. In order to simulate this phenomenon, ARTIST generates two types of sound sources: the *voicing source*, which produces a quasi-periodic vibration, and the *noise source*, which produces turbulence. The former generates a pulse stream intended to simulate the nonuniform (or quasi-periodic) vibration of the vocal folds, whereas the latter is intended to simulate an airflow past a constriction or past a relatively wide

https://doi.org/10.1017/S1355771898003070 Published online by Cambridge University Press

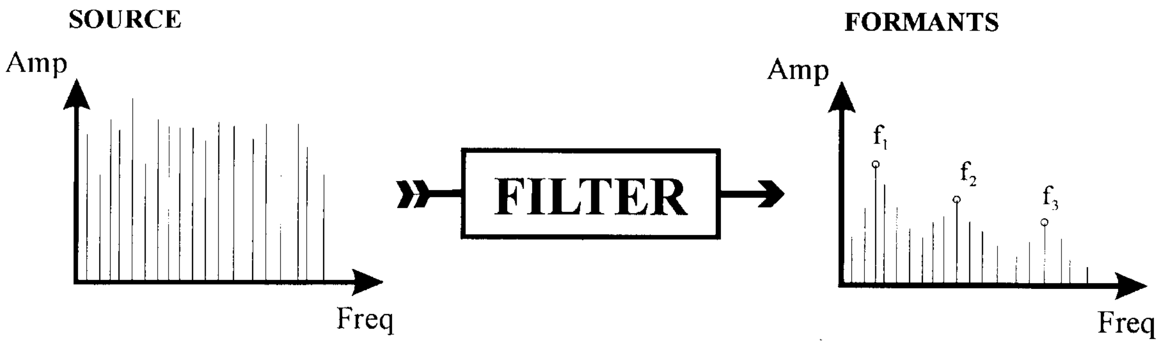

Figure 2. The subtractive synthesis approach to speech synthesis.

separation of the vocal folds (figure 3). Noise in the human voice corresponds to the generation of turbulence by the airflow past a constriction and/or past a relatively wide separation of the vocal folds.

The heart of the voicing source is a pulse generator (on the right side of the architecture shown in figure 3). It produces a periodic waveform at a specific frequency with a great deal of energy in the harmonics. A pulse waveform has significant amplitude only during relatively brief time intervals, called the pulse width. A pulse waveform has a rich spectrum. The output of the voicing source provides the raw material from which the filtering system will shape the required sound.

On its journey through the vocal tract, the sound is transformed. Components which are close to one of the resonant frequencies of the tract are transmitted with high amplitude, while those which lie far from a resonant frequency are suppressed. Much of the art of the singer lies in shaping the vocal tract in such a way that the crude source output is transformed into a desired sound. The vocal tract can be thought of as a pipe from the vocal folds to the lips plus a side branch leading to the nose, with a cross-sectional area which changes considerably.

The length and shape of a particular human vocal tract determine the resonance in the spectrum of the voice signal. The length of the human vocal pipe is typically about 17 cm, which can be slightly varied by raising or lowering the larynx and shaping the lips. The cross-sectional area of the tube is determined by the placement of the lips, jaw, tongue and velum. For the most part, however, the resonance in the vocal tract is tuned by changing its cross-sectional area at one or more points. A variety of sounds may be obtained by adjusting the shape of the vocal tract during phonation.

Five BPFs are appropriate for simulating a vocal tract with a length of about 17 cm; in figure 4, these correspond to the five first BPFs from left to right (the sixth BPF is used to simulate a nasal side branch). Each filter introduces a peak in the magnitude spectra determined by the pass-band centre frequency and by the formant bandwidth. The pass-band centre frequencies and bandwidths of the lowest three formants vary substantially with changes in articulation, whereas the fourth and fifth formants do not vary as much. Figure 4 portrays the basic architecture of the synthesis engine of ARTIST; more details about its functioning and an implementation

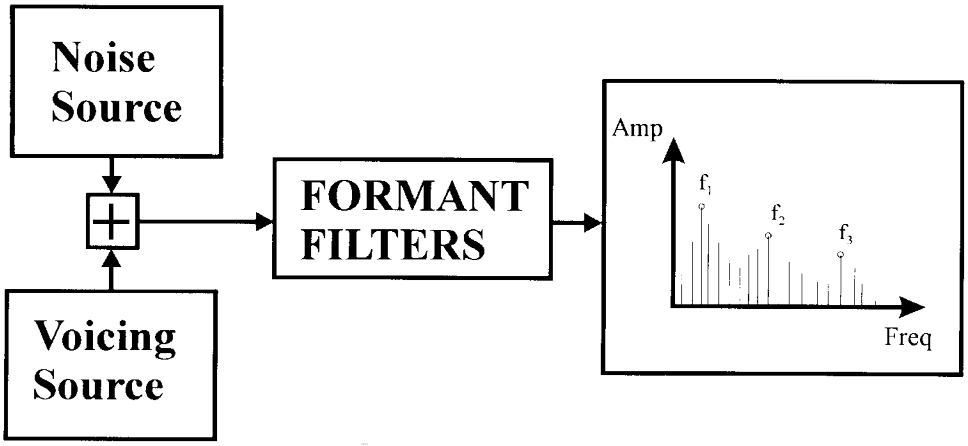

Figure 3. ARTIST generates two types of sound sources: the voicing source, which produces a quasi-periodic vibration, and the noise source, which produces turbulence.

https://doi.org/10.1017/S1355771898003070 Published online by Cambridge University Press

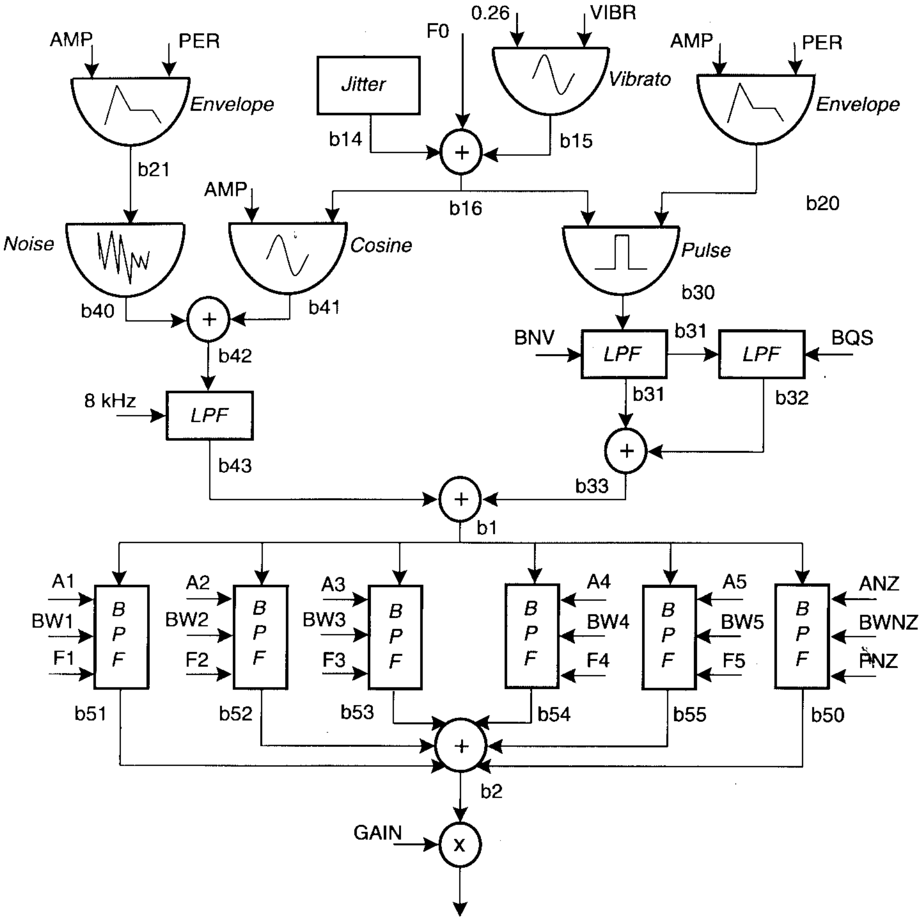

Figure 4. The basic architecture of a subtractive synthesizer.

in *pcmusic* (a programming language for sound synthesis) can be found in the book *Computer Sound Synthesis for the Electronic Musician* (Miranda 1998a).

3. THE MEZZO-SOPRANO PART

It has been demonstrated in section 2 that the spectral contour of vocal-like sounds has the appearance of a pattern of ‘hills and valleys’ (figure 2). The central frequencies (for example, f_1, f_2 and f_3, in figure 1), the amplitudes and the bandwidths of the peaks define the colour (or timbre) of a vocal sound. For example, the central frequency of the first formant of a vowel /a/ (as in the word ‘sanctus’ in Latin), sung by a tenor is approximately 331.12 Hz, whereas the value of the first formant of a vowel /o/ (as in the word ‘eleison’ in the Latin setting of the mass) is approximately 196 Hz.

Each vowel sound is associated with a specific formant configuration and by changing the shape of our vocal tract we change the lower formants as we speak or sing. The three lowest formants are the most important in making vowels recognisable. In male adults, the first formant can vary between 250 Hz and 1 kHz, the second can vary between 600 Hz and 2.5 kHz and the third between 1.7 kHz and 3.5 kHz.

The author is particularly interested in composing with vowel-like sounds, as he believes that our ability to recognise vowels is closely related to our ability to recognise timbre. He finds that this phenomenon is a good starting point for the definition of a framework to compose a piece.

The pitches and, to a certain extent, the whole harmonic structure of the Requiem were defined according to the values of the first three formants’ central frequencies for the vowels /a/, /e/, /i/, /o/ and /u/, as they are pronounced in Latin, sung by male and female voices. After much research and experimentation, the author defined a note-based formant system by matching the frequencies of musical notes to formant values, based upon the standard Western twelve-tone equal temperament system, with a tuning reference of A4 = 440 Hz (figure 5).

These notes correspond to formant centre frequency values of five vowels (/a/, /e/, /i/, /o/, /u/) sung by male (M) and female (F) synthesised voices. From the top stave down, the notes correspond to the third, second and first formant centre frequencies. The notes on the bottom stave are one octave below their respective first formant values.

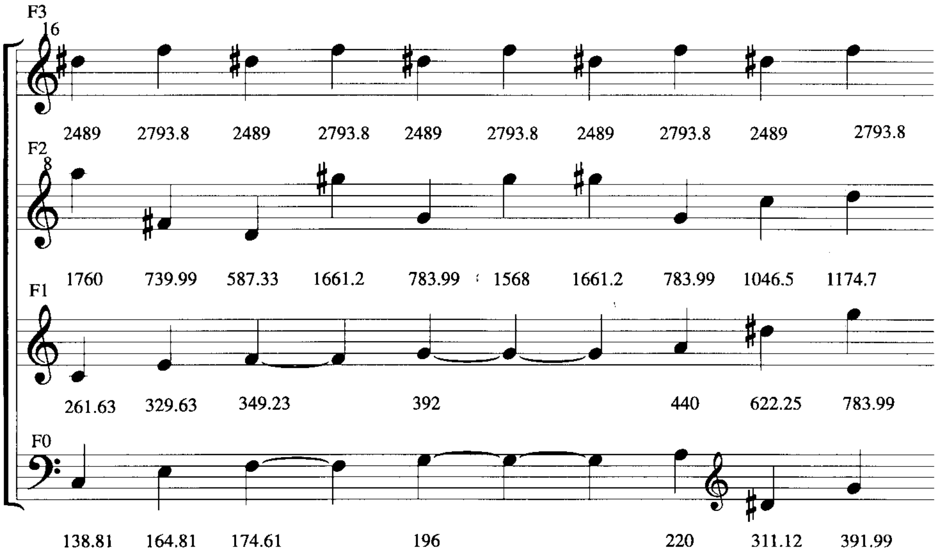

Figure 5. The pitch material for the lyrics of the Requiem was derived from formant centre frequency values of five vowels (/a/, /e/, /i/, /o/, /u/) sung by male (M) and female (F) voices.

These notes were organised into a number of different pitch-sets (e.g. F0, F1, F2 and F3 sets, a set of vowel /a/ values, etc.). Then, the individual sets were used to produce the musical passages for the mezzo-soprano part. Figure 6 illustrates a typical excerpt from the *Kyrie* using notes from the F0 pitch-set.

It is important to mention that the use of formant values to derive compositional material is by no means new. Composers such as Karlheinz Stockhausen in his fabulous *Stimmung* (Maconie 1976) and Wayne Slawson in *Quatrain 2* (Slawson 1987) are the two most typical examples that come to mind, but it was probably Hermann Helmholtz who established the theoretical foundations for such a composition framework in his classic book *Die Lehre von den Tonempfindungen* (On the Sensations of Tone), first published in the 1870s (Helmholtz 1954). What is perhaps different in the approach outlined here is that such systematisation was used to forge the integration of electroacoustically generated material and the (natural) human voice.

4. THE SCORE

The author opted for the design of a very straightforard layout for the score. He is of the opinion that there is no need to invent strange symbols to convey information that can be conveyed by well-established, popular means. The mezzo-soprano part of the Requiem was written as clearly as possible, using traditional notation and only a few nonstandard symbols were created to indicate when an effect should take place or whether or not the electroacoustic material is playing. Figure 7 illustrates an excerpt from the score. The symbol on the top of the stave at the beginning of the second bar indicates that effect number 2 should be activated at that point (an index of effects is supplied with the score) and below the stave there is an indication for the entrance of this particular passage. This indication gives timing information and a brief description of what is recorded on the tape. The performer has a certain degree of freedom to establish when she will start to sing. So far the piece has two variations by two different performers (the version on the CD features Beirut-born soprano Simone Sahyouni), and both have established their own aural cues during the rehearsals.

5. CONCLUSION

This paper has outlined some compositional aspects of *Requiem per una veu perduda*. It has discussed one of the synthesis techniques used to synthesise speech and demonstrated the method for the specification of the basic compositional material. It is proposed that

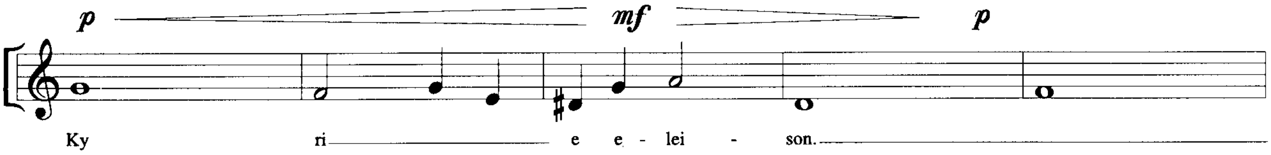

Figure 6. A typical passage from the *Kyrie*.

https://doi.org/10.1017/S1355771898003070 Published online by Cambridge University Press

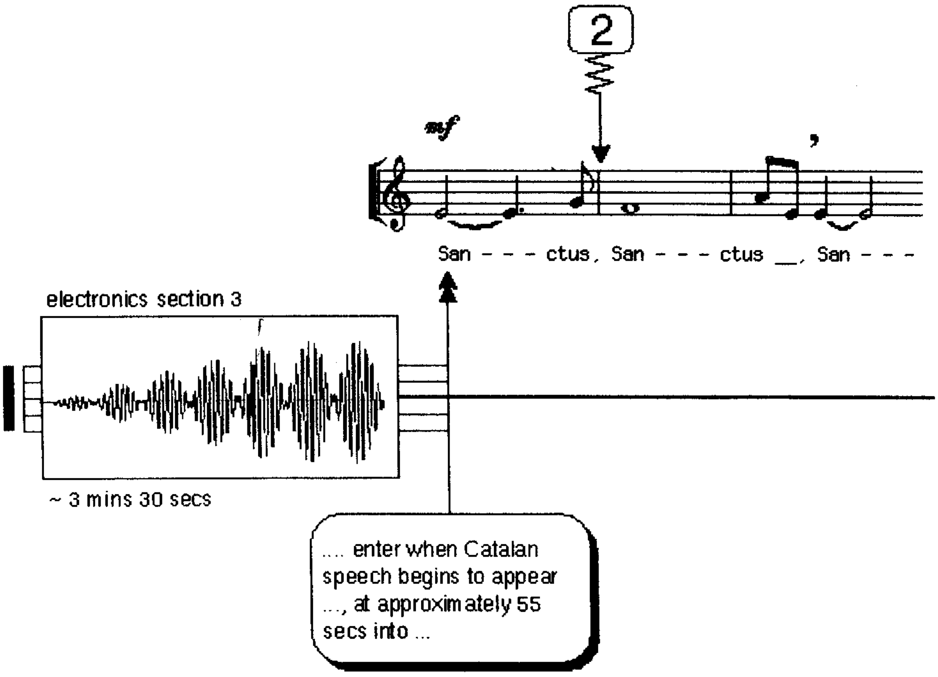
2
mf
San - - - ctus, San - - - ctus __, San - - -
electronics section 3
~ 3 mins 30 secs
.... enter when Catalan speech begins to appear ..., at approximately 55 secs into ...

Figure 7. An excerpt from the score.

the close relationship between the synthesis techniques used to produce most of the electroacoustic materials, and the pitch system derived from the inner structure of vocal sounds, plays a key role in the coherence of the composition as a whole.

REFERENCES

Arbo, A. 1998. La truce du son. Expression et intervalle chez Condillac. In J.-M. Chouvel and M. Solomos (eds.) *L'espace: musique/philosophie*. Paris: Editions L'Harmattan.

Fishman, R. 1997. The phase vocoder: theory and practice. *Organised Sound* **2**(2): 127–45.

Helmholtz, H. L. F. 1954. *On the Sensations of Tone*. New York: Dover Publications.

Maconie, R. 1976. *The Works of Stockhausen*. London: Marion Boyars.

Miranda, E. R. 1995a. Chaosynth: Um sistema que utiliza um automato celular para sintetizar particulas sonicas. Porto Alegre: *Proc. of II Brazilian Symp. on Computer Music*, pp. 205–12.

Miranda, E. R. 1995b. An artificial intelligence approach to sound design. *Computer Music Journal* **19**(2): 59–75.

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

Miranda, E. R. 1998b. Striking the right note with ARTIST: an AI-based synthesiser. In M. Chemilier and F. Pachet (eds.) *Recherchers et applications en informatique musicale*, pp. 227–39. Paris: Hermes.

Serra, X. 1997. Musical sound modeling with sinusoids plus noise. In C. Roads *et al.* (eds.) *Musical Signal Processing*, pp. 91–122. Lisse: Sweets and Zeitlinger.

Slawson, W. 1987. Sound–Color Dynamics. *Perspectives of New Music* **25**(1–2): 156–77.

Thomas, D. A. 1995. *Music and the Origins of Language*, Cambridge, UK: Cambridge University Press.

https://doi.org/10.1017/S1355771898003070 Published online by Cambridge University Press