THE DEVELOPMENT OF A SINGING SYNTHESISER.

Ken Lomax
OxfordUniversity and IRCAM, Paris.
c92kl@comlab.ox.ac.uk   http://www.phon.ox.ac.uk/~lomax/homepage.html

Abstract

This paper describes the development in software of a singing synthesiser which takes a musical manuscript consisting of lyrical and melodic data and synthesises the vocal part in one of several voices. Work has centred around two domains: the design of a synthesis controller and the creation of a synthesis engine. The synthesis controller consists of two steps: the first interprets the manuscript into a set of musical intentions and the second realises these as acoustical control parameters. The synthesis engine is based on the manipulation and concatenation of sound-units, typically diphones using spectral modelling synthesis[Serra, 1990; Rodet, 1991].

1 Introduction

This paper describes the development in software of a singing synthesiser which takes a manuscript consisting of lyrical and melodic data and synthesises the vocal part in one of several voices and styles. Work has centred around two domains: the design of a synthesis controller and the creation of a synthesis engine. The synthesis controller consists of two steps: the first interprets the manuscript into a set of musical intentions and the second realises these as acoustical control parameters. The synthesis engine manipulates and concatenates sound-units, typically diphones using spectral modelling synthesis [Serra, 1990; Rodet, 1991]. The architecture of the synthesiser is illustrated in figure 1.

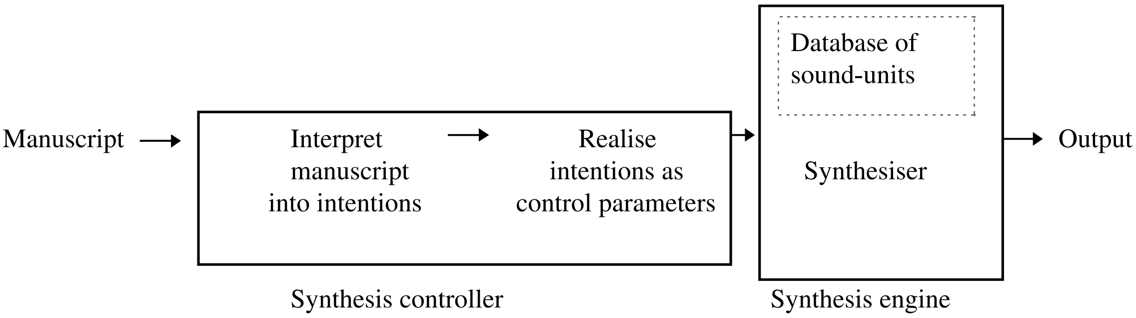


Figure 1: The architecture of the singing synthesiser.

In this paper the manuscript creator, synthesis engine and synthesis controller are reviewed in turn.

2 The manuscript creator

The manuscript must encompass the necessary melodic and lyrical data and is created using a MIDI keyboard and graphical interface (figure 2).

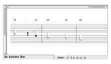


Figure 2: The manuscript creator.

The lyrics currently take the form of a series of sound-units with additional characters such as '_' and '-' aligning the units to the melody. Once the melodic and lyrical data have been defined the synthesis controller may be called.

3 The synthesis controller

The role of the synthesis controller is to interpret from the manuscript a set of dynamic acoustical parameters with which to drive the synthesis engine. At KTH, a set of performance rules have been produced which introduce deviations to the notated tempo, amplitude and pitch [Carlson, 1988; Friberg, 1995]. These have been developed to reflect the process of musical interpretation and performance characteristics of a singer [Krom, 1995; Cook, 1995;Prame, 1994: Sundberg, 1987; Seashore, 1938]. The derivation of these rules was based upon analysis by synthesis techniques in which a musical phrase was played with and without a rule, and musicians asked to comment on any perceived improvement. The task was therefore approached from the perspective of the observer. In this paper a complimentary approach is taken which attempts to view the act of singing from the perspective of the singer himself. An attempt is made to rationalise and simulate the mental process underlying singing and is based on the following propositions: 1) *when a singer reads a musical manuscript a mental process of interpretation is applied which translates the notation into a set of musical intentions, 2) the singer performs these intentions resulting in the singing voice.*

3.1 From a manuscript to intentions.

It is proposed that when a singer reads a manuscript it is interpreted into a series of musical intentions. An intention might be for example 'Sing 'aa' to middle C' or 'Crescendo to ff' or 'End the note'. A set of possible intentions may be derived by asking singers which actions they would perform and in what order to sing a given phrase. Consider the example in figure 3.

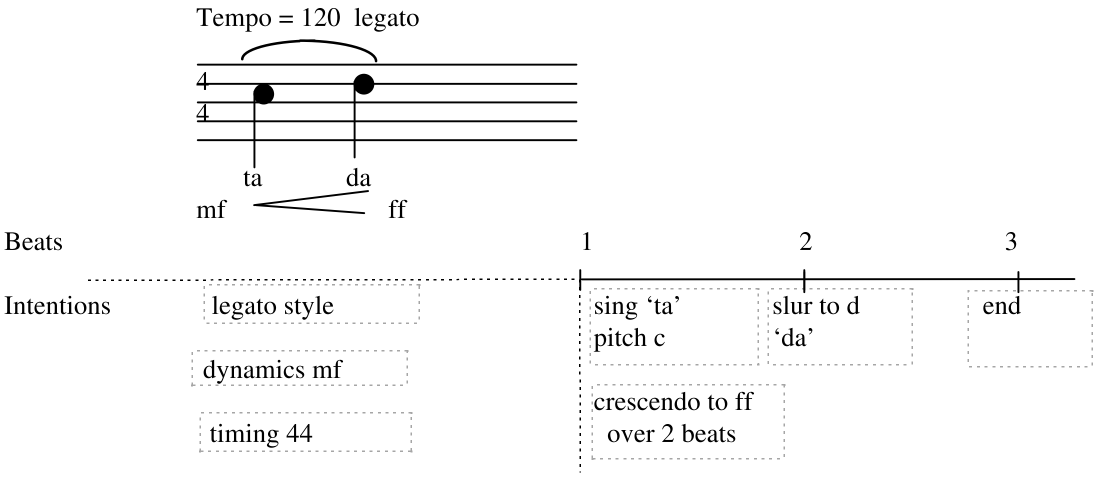


Figure 3: A musical phrase and resulting intentions.

The resulting intentions illustrate how one might attempt to sing the notated phrase. The singer intends to place or sing 'ta' on beat one, then to slur to the note d with the lyric 'da' while applying a crescendo to *ff,* and finally ending on beat three. The intentions to the left of beat one reflect the fact that singers are aware of global conditions taken from the manuscript. For example the phrase is to be sung in a $^{4}_{4}$ rhythm with an initial dynamic of *mf* and is to be sung in a legato style.

To simulate this process a manuscript parser is created consisting of a database of intentions and parsing rules which translate musical notation into musical intentions. Each intention contains a selection of parameters which stipulate the precise manner in which the intention is to be performed(figure 4).

Intention	Parameters
Sing a note	lyric, pitch, dynamic, duration, style
Slur to a note	lyric, pitch, dynamic, duration, style
Release	style
Crescendo	to dynamic, duration
Ritardo	to tempo, duration
Sing to 44 rhythm	

Figure 4: A table of intentions.

By applying the parsing rules to the manuscript a set of musical intentions can be derived sharing the following properties: 1) the set of intentions is mainly sequential, 2)some intentions apply to individual notes and others to whole phrases and 3) each intention may affect only those that follow. (figure 5).

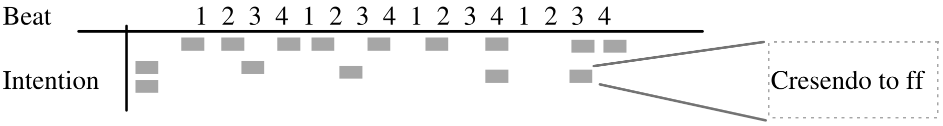


Figure 5: A set of intentions arising from a musical phrase.

3.2 Realising the intentions

It is proposed that for trained singers, the manner in which a musical intention is realised displays regularities. These conform to ideal notions which the singer has developed on how each intention should best be realised[Hewitt 1985; Seashore, 1938]. For example if a trained singer is asked to perform the intention 'Start singing 'aa' to C at mf legato' then characteristic envelopes will be observed(figure 6)

[Cook, 1995; Prame,1994; Krom, 1995; Sundberg, 1987; Friberg, 1995; Ternstrom *et al.* 1989; Castellango *et al.* 1989].

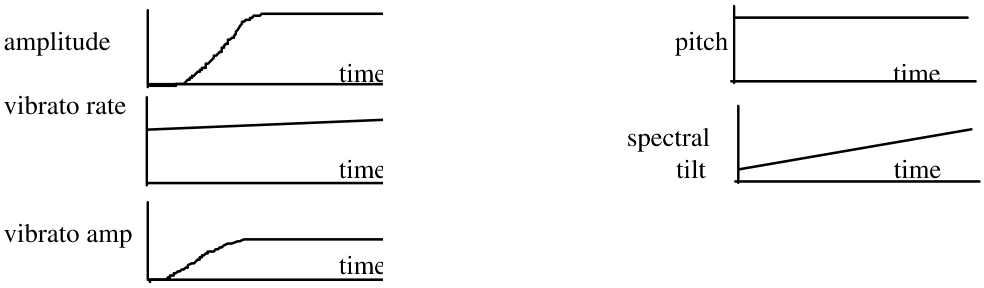
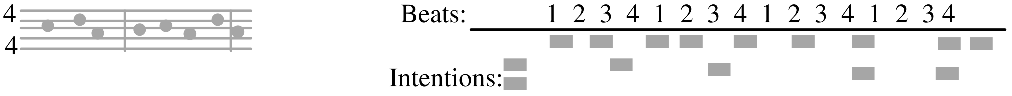


Figure 6: Characteristic performance of the intention Sing 'aa' to C at mf legato'

These reveal a typical amplitude attack with a gradual increase in vibrato depth and rate, a pitch path fluctuating around the correct frequency and a rise in spectral tilt.[Cook, 1995; Prame,1994; Krom, 1995; Sundberg, 1987; Friberg, 1995; Ternstrom *et al.* 1989; Castellango et al. 1989]. These envelopes may be encoded in a template which may be referenced by the synthesis controller when the intention 'Start singing' is observed. Similar templates may be created to descibe the realisation of many other musical intentions found in the performance of the singing voice. [Bresin and Vecchio 1995; Bresin and Ghetta, 1995; Gabrielsson, 1995; Sundberg, 1993; Sundberg, 1992; Sloboda, 1983; Castallengo, 1993; Cook, 1995]

The synthesis controller may now take each intention and derive from each, a set of parametric envelopes from appropriate templates. The resulting parameters are stored in graphs sharing a common time axis. This entire process is summarised in figure 7.



1) Manuscript

2) The manuscript is translated into intentions

3) The intentions are realised

4) This yields the acoustic parameters.

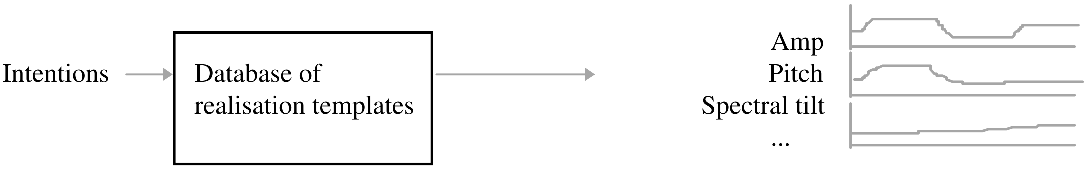


Figure 7: A summary of steps one and two.

Once the acoustical parameters have been derived the synthesis engine may be called.

4 The synthesis engine

Of the many speech synthesises techniques developed, three have been used extensively in singing. Work at KTH has concentrated on formant synthesis [Larsson, 1977; Klatt, 1980] articulatory models have been developed at CCRMA [Cooke, 1993 & 1991] and IRCAM have produced concatenative additive synthesisers [Rodet, 1991]. Other projects of note include the Chant Project [Rodet, 1985], frequency modulation synthesis [Chowning, 1979] and PSOLA synthesis [Dan *et al.* 1993; Moulines *et al.* 1990]. A compact disc recording featuring six synthesised voices each created with different systems provides an excellent review of the current quality of singing synthesis [Sundberg and Nordlindh, 1995]. In this project, vocal output of a high and personal quality is desired and while formant and articulatory synthesis may achieve this at times, the manipulation and concatenation of sampled sounds provides this directly. For this reason the synthesiser is based currently on the manipulation and concatenation of sound-units using spectral modelling synthesis [Serra, 1990; Rodet, 1991]. This operates on a database of sound-units which is created as follows.

4.1 Creating a database of sound-units.

The sound-units which are used currently are derived from the nonsense word /tʌɾʌmiʌduʌlʌiu/ or "tadameadoalaeo". This has been chosen as it contains a number of sounds appropriate for scatting; that is singing with a restricted number of sounds. Each singer is asked to sing this word at various pitches and dynamics from which the sound-units are extracted, analysed and encoded. The analysis routines return the formant trajectories, a waterfall of continuous spectral envelopes [Serra, 1990], spectral tilt, dynamic and pitch paths. This data is stored in a sound-unit which may contain other renditions of the same sound sung at other pitches and dynamics. This process is summarised in figure 8.

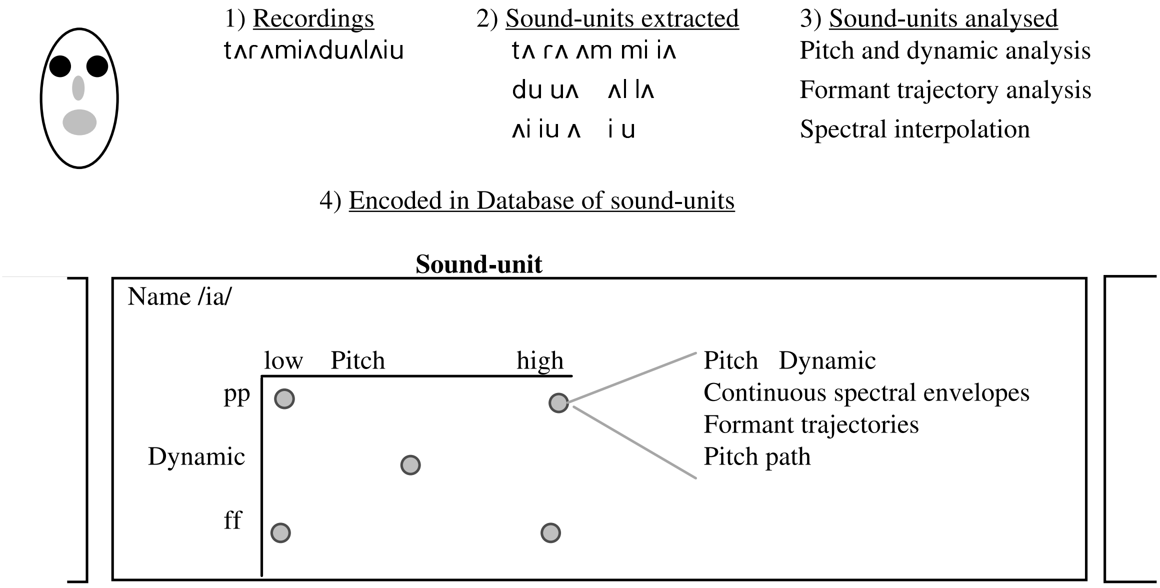
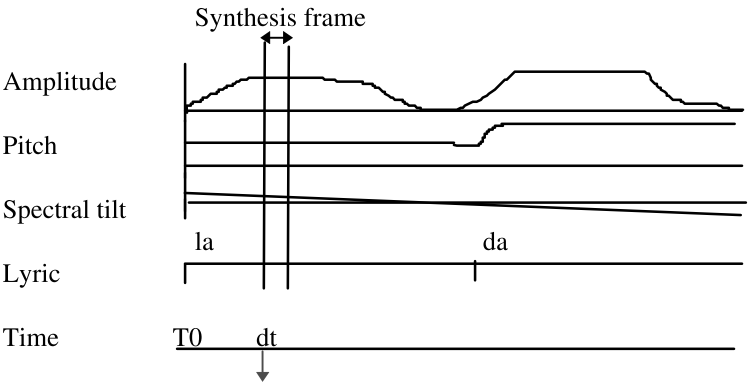


Figure 8: Encoding a sound-unit in a database.

The sound-units are now in a format which allows their resynthesis at different pitches and durations using additive synthesis.

4.2 Creating the output with additive synthesis.

The synthesis engine is supplied with a set of graphs depicting the amplitude, pitch, lyric and spectral tilt parameters against time. The role of the synthesis engine is to iterate along these graphs at intervals defined by a synthesis frame and at each point to sample the amplitude A, pitch P, spectral tilt S and the current lyric (figure 9).



Pitch P, Amplitude A, Spectral tilt S, Lyric ‘la’.

Figure 9: Querying the parameters at time t.

With this information a spectral envelope may be extracted from the database. In this instance the sound-unit ‘la’ which was sung closest to pitch P is located and the spectral envelope E returned, where E lies at (dt-T0) along the waterfall of spectra in the sound-unit. It is now possible to generate a set of partials as illustrated in figure 10.

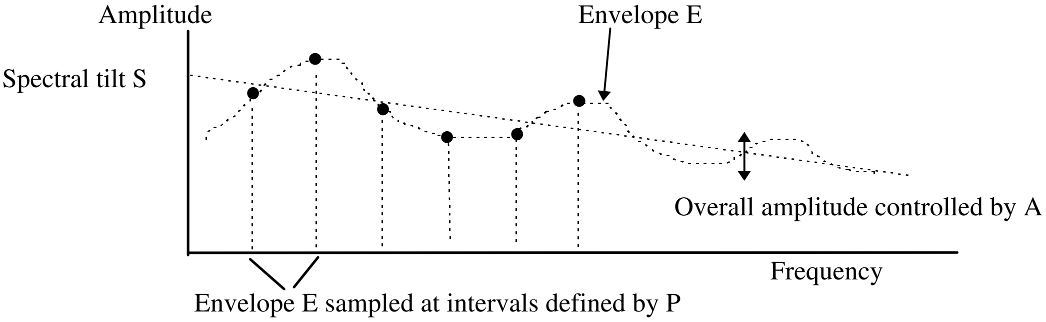


Figure 10: A set of partials may be derived from an envelope E, pitch P, amplitude A and spectral tilt S.

The set of partials can then be used to synthesise the wave for the duration of the synthesis frame using additive synthesis[Serra, 1995, Freed *et al.* 1993]. This entire process iterates along the time axis for the duration of the musical piece resulting in the complete synthesised vocal part.

5 Future work

Future work should concentrate on the enhancement of the synthesis controller by improving the realisation routines. For this purpose, several opera singers have been asked to sing notated phrases from which intentions and their associated realisations may be derived. For each intention, any influencing parameters must be noted and catered for. For example the intentions ‘Start singing “aa” to middle C’ will be affected by the style marked, the position in the bar and dynamic markings. The way in which these parameters influence the performance or realisation of the intentions must be encoded. Additional rules controlling for example small scale pitch deviations and the timing of pitch transitions must also be encoded. These may be developed from rules at KTH.

6 Conclusion

In this project, research has followed two paths: firstly to develop a synthesis engine and secondly to create a controller. The synthesis engine is successfully manipulating sound-units and the control system is producing musical phrases. Work will focus currently on encoding the realisation of musical intentions for several classical singers so that sung phrases may be synthesised not only with the voice of a specified singer but also in their style. The addition of musical ornaments such as trills and turns will be considered.

Acknowledgements

I am most grateful to Xavier Rodet and his team at IRCAM for allowing me to work with them for three months last year. Over this time I gained valuable insight into concatenative di-phone synthesis and spectral modelling synthesis. I am grateful too to my supervisors Dr. John Coleman and Ian Page for their input.

References

[Berndtsson, G. 1995] *Systems for synthesising singing and for enhancing the acoustics of music rooms.* Stockholm: Speech Communication and Music Acoustics, KTH.

[Bresin, R., Poli, G. and Ghetta, R. 1995] Fuzzy performance rules. In A. Friberg and J. Sundberg (eds)*Grammars for music performance*. Stockholm: Speech Communication and Music Acoustics, KTH, 15-36.

[Bresin, R. and Vecchio, C. 1995] Neural networks play Shumann. Analysis and synthesis of the performing action of a real pianist. In A. Friberg and J. Sundberg (eds) *Grammars for music performance*. Stockholm: Speech Communication and Music Acoustics, KTH, 5-14.

[Carlson, R. 1988] *The KTH program for synthesis of singing*. Stockholm: Speech Communication and Music Acoustics, KTH.

[Castallengo, A. 1993] Fusion or separation: from vibrato to vocal trill. In A.Friberg and J. Iwarsson (eds)

Proceedings of the Stockholm Music Acoustics Conference. Stockholm: Speech Communication and Music Acoustics, KTH. 79-82

[Castellengo, M and Richard, G. and Alessandro, C. 1989] Study of vocal pitch vibrato perception using synthesis. In 13th International congress on acoustics. Yugoslavia.

[Chowning, J.M. 1979] Computer synthesis of the singing voice. Stanford: CCRMA, Stanford University.

[Cook, P.R. 1995] An investigation of singer pitch deviation as a function of pitch and dynamics. In [K.Elenius and P.Brabderund (eds*) Proceedings of the XIIIth international congress of phonetic sciences*. Stockholm: Speech Communication and Music Acoustics, KTH. 202-205.

[Cook, P.R. 1993] *SPASM, a Real-Time Vocal Tract Physical Model Controller; and Singer, the Companion Software Synthesis System*. Stanford: CCRMA, Stanford University.

[Cook, P.R. 1991]. *LECTOR: An Ecclesiastical Latin Control Language for the SPASM/singer instrument*. Stanford: CCRMA, Stanford University.

[Dan, T. and Mukherjee, B. and Datta, A.K. 1993] Temporal approach for synthesis of singing. In A.Friberg and J. Iwarsson (eds) *Proceedings of the Stockholm Music Acoustics Conference*. Stockholm: Speech Communication and Music Acoustics, KTH. 282-286.

[Freed, A. and Rodet, X. and Dephalle, P. 1993] Synthesis and control of hundreds of sinusoidal partials on a desktop computer without custom hardware. Paris: IRCAM.

[Friberg, A. 1995] *A Quantative Rule System for Musical Performance.* Stockholm: Department of Speech Communication and Music Acoustics, KTH.

[Gabrielsson, A. 1995] *Music Performance*. Uppsala: Department of psychology, Uppsala University.

[Hewitt, G. 1985] How to sing. Cambridge: EMI Music.

[Klatt,D.H. 1980] Software for a cascade/parallel formant synthesiser. In *Journal of the acoustical society of America*. Vol.67. 1971-1995.

[Krom, G. 1995] Timing and accuracy of fundamental frequency changes in singing. In K.Elenius and P.Brabderund (eds) *Proceedings of the XIIIth international congress of phonetic sciences*. Stockholm: Speech Communication and Music Acoustics, KTH. 205-221.

[Larsson, B. 1977]. *MUSSE - Music and Singing Synthesis Equipment.* Stockholm: Department of Speech Communication and Music Acoustics, KTH.

[Moulines, E. and Charpentier, F. 1990] Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Speech Communication.* Holland:Elsevier Science. Vol. 9. 453-467.

[Prame, E. 1994] Measurements of the vibrato rate of ten singers. *In Journal of the acoustical society of America*, 1979-1984.

[Rodet, X. 1985] *Chant Manual*. Paris: IRCAM.

[Rodet, X. 1991] *Generalized diphone control*. Paris: IRCAM.

[Seashore, C.E. 1938] *Psychology of music.* New York: Dover Publications.

[Serra, X. 1990] *Spectral Modelling Synthesis: A Sound Analysis/Synthesis System based on a Deterministic plus Stochastic Decomposition.* Stanford: Stanford University.

[Sloboda, J.A. 1983] Music Performance. In *The Psychology of music.* New York: Academic Press Inc.

[Sundberg, J. 1987] *The science of the singing voice*. Illinois: University Press.

[Sundberg, J. 1982] Speech, song and emotions. In M.Clynes (ed) *Music, mind and brain*. England: Plenum publishing corporation.

[Sundberg, J. and Nordlindh, T. 1995] *Information technology and music CD-ROM.* Stockholm: Speech Communication and Music Acoustics, KTH

[Sundberg, J. 1993] How can music be expressive? In C. Sorin (ed) *Speech Communication 13* Holland:Elsevier Science 239-253.

[Ternstrom, S. and Friberg, A. 1989]. Analysis and simulation of small variations in the fundamental frequency of sustained vowels. In *Speech Transmission Laboratory Quarterly Report.* Stockholm: Speech Communication and Music Acoustics, KTH. 282-286.