RESEARCH ASPECTS ON SINGING

AUTOPERCEPTION
COMPUTER SYNTHESIS
EMOTION
HEALTH
VOICE SOURCE

Papers given at a seminar organized
by the Committee for the Acoustics of Music

Publications issued by the Royal Swedish Academy
of Music No. 33
1981

SINGING SYNTHESIS IN ELECTRONIC MUSIC

By GERALD BENNETT, Professor of composition, Musikhochschule Zurich, Schweiz

It is a special honor for me as a musician to have been invited to a scientific symposium on the voice as a musical instrument. My work on the singing voice has not shared the objective goals of that of my fellow speakers today. While they have primarily been concerned with the understanding of complex physiological, acoustical, and linguistic phenomena for knowledge's sake, I have tried to answer certain acoustical questions in order to make music using the computer.

I will present the results of a study of four soprano voices. The purpose of the study was to improve a program for sound synthesis by computer which had as model the human vocal tract. I will speak principally about the results of this research; at the end of the talk, however, I shall try to indicate the importance of this and similar research for computer sound synthesis in particular and electronic music in general.

Before I begin, I want to acknowledge two debts. The first is to Gunnar Fant, whose theory of speech production is of central importance for this project. The second is to Johan Sundberg, without whose wise and precious assistance this project could never have begun. Neither of these debts can be repaid; I am proud to be able to speak of the fruits of their inspiration and assistance in the presence of both men.

In December, 1978, at the Institut de Recherche et Coordination Acoustique/Musique (IRCAM) in Paris, Xavier Rodet, Johan Sundberg and I began to synthesize singing voices using a computer, not with any of the available languages for computer sound synthesis, but rather by means of a very powerful and efficient program for speech synthesis developed by Rodet some years earlier at the laboratories of the Commission à l'Energie

Atomique (CEA) at Saclay, France. Although the study I will speak of today is my own, the vocal synthesis project at IRCAM has been carried out by Rodet and me jointly; without his sound synthesis program, none of this research would have been done.

Details of the techniques of synthesis Rodet uses can be found in Rodet & Bennett (1980). Briefly, his program models the resonant response of an imaginary vocal tract to a virtual source, not as is usually done, by filtering the source to obtain the resonant peaks, or formants, but rather by synthesizing directly the sound of each formant and then combining the formants to a single voice. This method allows the user to define and control the resulting sound with great precision. Other seminars in music acoustics at Stockholm have dwelt at length on techniques of sound production and reproduction by computer. I shall not recapitulate any of these discussions here, except to direct the curious reader to the standard text Mathews (1970).

To begin with, listen to two examples of singing voices synthesized on the computer using Rodet's method. The first (SOUND EXAMPLE 1) is a phrase from the madrigal "Moro lasso" by Carlo Gesualdo (ca. 1560-1613). The second (SOUND EXAMPLE 2) is a composite of computer voices I used in a piece for baritone, five instruments and tape: Aber die Namen der seltnen Orte und alles Schöne hatt' er behalten

Rodet and I were happy that we had been able to synthesize sounds bearing at least some resemblance to natural singing voices, but we found our efforts disappointing from many points of view. The most glaring fault seemed to be the lack of individuality of the voices, and after I had finished the piece for baritone, instruments and tape in November, 1979, I set out to remedy this failing.

By this time, after a year's work, we had learned the following things: We knew how to synthesize various sung vowels by defining formant frequencies, amplitudes and bandwidths; we could distinguish male and female voices in synthesis by differences in formant position for the same vowel, by differences in spectral richness, and by differences in the behavior of the formants with changes of fundamental frequency; we were able to link perceived amplitude and spectral richness in a convincing way (so in SOUND EXAMPLE 2 every crescendo and decrescendo brings about a change of timbre in the voice); we understod something of the nature of both vibrato and the apparently random fluctuations of fundamental frequency we observed in every singing voice; we modified an algorithm of Fant's defining formant amplitude and bandwidth in terms of formant frequency so as to simplify greatly the use of the synthesis program. But we always synthesized the same soprano or the same baritone; we did not know how to define several voices of the same register, each having its own "personality". The literature offered only average values for acoustical parameters - those we had used at the beginning of our synthesis - but gave no clue as to how these values vary among similar voices. I was looking for specific values for the parameters which are given to the synthesis program: formant frequencies, vibrato rate, etc; since most of our work hitherto had been with male voices, I decided to find these values by analysing carefully four different soprano voices, three of professional women singers and one of a boy soprano, aged 11 years.

I made extensive recordings (two to three hours each) of the three female voices, less extensive ones for the boy's voice. I asked the singers to sing single notes on specific vowels, at varying subjective dynamic levels throughout their entire ranges. In addition I asked special studies of each of the women (for example many different attacks on one pitch, different types of vibrato on the same note, etc). The recordings provided the material for the present study.

My first and certainly most important discovery was that specific single values for the parameters I sought to define were inadequate to characterize a specific voice. No one vibrato rate would distinguish voice A from voice B; much more important turned out to be the way in which the vibrato rate of one voice varied under specific conditions, compared to that of another voice. I very soon realized that I was looking not for unique values but rather for rules of behavior of the most important acoustical features for each voice. A first analysis suggested that the following five aspects of each voice should be studied carefully:

1. timbre (spectrum)
2. overall sound pressure level
3. vibrato rate and amplitude
4. random variation of fundamental frequency
5. attack patterns

I shall speak about each of these five aspects separately.

1. Timbre

For each of the four voices I tried to model the acoustical spectrum as precisely as possible. In our synthesis program this meant simply defining center frequencies for each formant, for the program calculated the appropriate formant amplitudes and bandwidths itself. In fact, this automatic calculation of formant bandwidth and amplitude worked well for the women's voices; it did not work at all for the boy's voice, presumably because our algorithm was derived from measurements of adult voices. Hence, SOUND EXAMPLE 3 gives only examples of synthesis of the women's voices. In the sound example one can compare the timbres of the natural voices with those of the synthesized voices; in the example, a short section of a sung note is followed by the synthesized timbre for each of the three singers.

Fig. 1a and 1b show the spectrum of one of these sung notes and that of my imitation respectively. Frequency is on the horizontal axis and ampli-

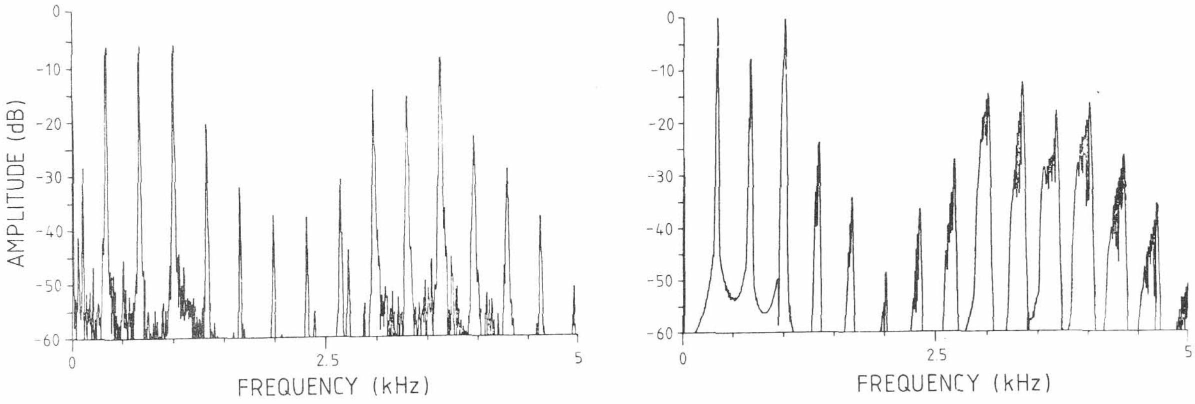

Fig. 1.

tude of the vertical, measured in dB less than the maximal amplitude found in the spectrum. The regularly spaced rays represent the sound's partials.

You will remark that in the lower part of the spectrum the imitation is quite good, each partial having nearly the amplitude of the corresponding partial in the original. In the higher end of the spectrum the imitation is less exact: this synthesized sound shows somewhat more energy in the upper formants. This imprecision may be due to the inexactness of the automatic calculation of amplitude and bandwidth of each formant, or it may be due to a difference in the source spectrum slopes of the real and the synthetic voices (Sundberg 1981). In the following discussion we shall see how to correct this difference.

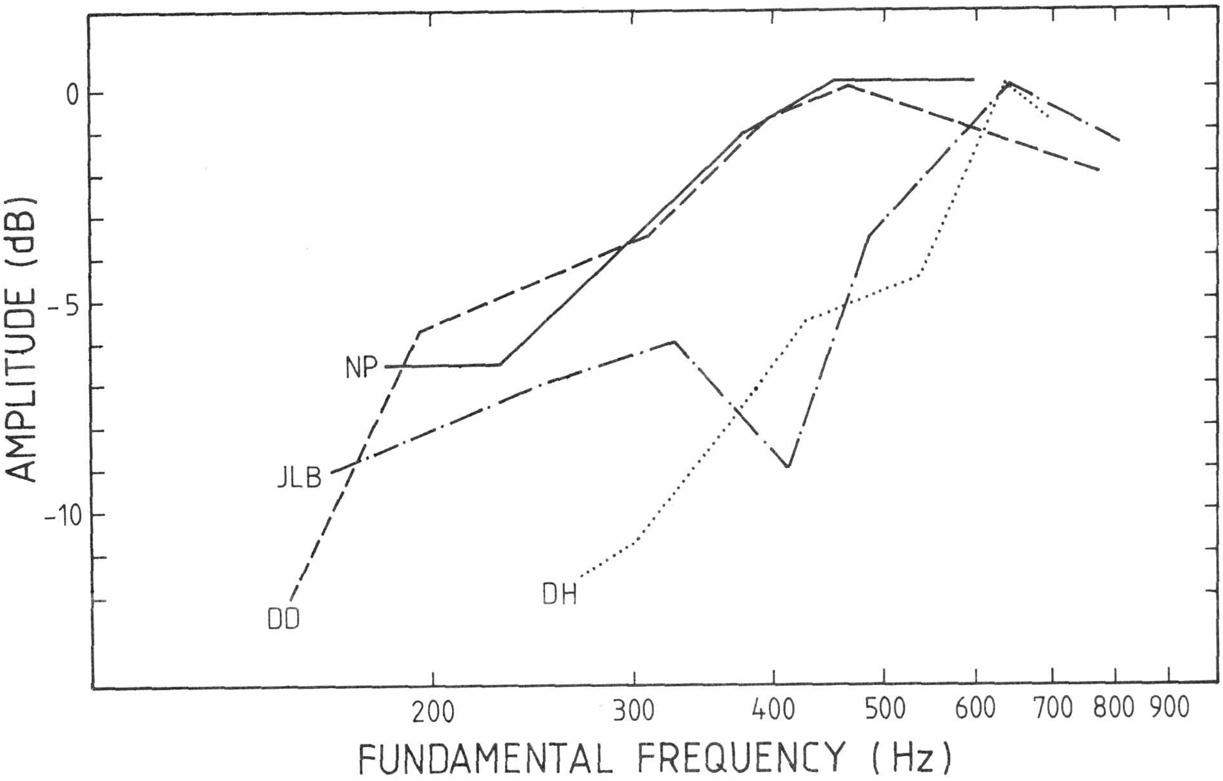

Fig. 2

2. Overall sound pressure level

I asked each singer to perform arpeggios throughout her or his range, keeping the sound pressure level (SPL) as nearly equal as possible. Each of the four singers sang more loudly in the upper part of her or his range than in the lower, but for each of the singers the SPL increased amplitude at a different rate and reached its maximum at a different point in the range. Not the specific values for any one pitch, but rather the behavior of the SPL over the entire range was characteristic and individual for each singer. Fig. 2 illustrates the results of the measurements of overall SPL as a function of pitch.

Here the change in amplitude clearly seems to be a function of fundamental frequency. The vertical scale shows how much softer than the loudest note each note of the arpeggio was. You see for instance that singer DD had the greatest dynamic change (but also the greatest pitch range), singer NP the least. You also see that JLB has a "break" in her voice around 415 Hz (G#4). Singer DH, the boy soprano, has a less dramatic break around 550 Hz (C#5). In my imitations, I took account of the slope of the increasing SPL, the frequency at which it was at a maximum and the slope of its decrease.*

When a singer makes a crescendo, he changes not only the loudnesse of the note sung, but also its timbre. In particular, the amplitude of the upper formants increases more rapidly than that of the fundamental as the note gets louder. This change in spectrum reflects change in the voice source due to the increased subglottic pressure. We had already modelled this timbral change in our original synthesis program, thanks to data given us by Johan Sundberg. We had originally assumed that formants two and higher

* These measurements of SPL are very approximate. For more precise measurements, microphone position and distance would need to be controlled more exactly. Since the goal of these studies was to trace patterns and not to define unique values, this degree of precision seemed unnecessary, especially in view of the much more important amplitude changes singers constantly make for expressive reasons.

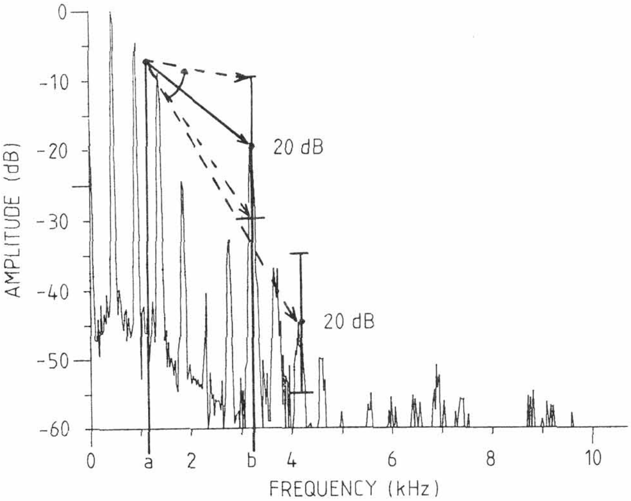

Fig. 3.

all changed amplitude by the same amount. Later we felt it more appropriate to increase the amplitude of the higher formants faster than that of the lower formants. Fig. 3 shows the principle of the present algorithm.

In analysing my recordings, I found the amplitude of the formant "hill" around the third and fourth formants of the loudest notes to be about 20 dB greater than in the softest notes (in the illustrations of spectra all amplitudes are scaled to the loudest partial, usually the fundamental, but sometimes - see Fig. 1a - the second partial). In our synthesis we now model this 20 dB change, linking it approximately monotonically to the overall SPL of the note. In addition, we can differentiate individual voices by defining a) the highest frequency whose amplitude does not increase faster than the overall SPL, and b) the frequency to which the observed 20 dB change applies. These points are marked a and b respectively in Fig. 3. Intermediate amplitudes are scaled proportionally. Our synthesis allows no direct control over the spectrum of the source assumed to be exciting the formants. Therefore we must mimic the effects of spectral change in the source in the resulting sound itself. The algorithm which does this is designed so that the amplitudes of all formants below frequency b vary with overall SPL less than 20 dB, while the amplitudes of the formants above frequency b vary more than 20 dB. The lower the frequency of b, the greater the increase in amplitude of the higher formants. The result is roughly equivalent to varying the slope of the spectrum of the voice source with the overall SPL. In SOUND EXAMPLE 4 you hear scales sung by synthetic voices using the timbral models of the three woman singers and imitating the patterns of SPL evolution shown in Fig. 3 (without the break in JLB's voice).

3. Vibrato rate and amplitude

Vibrato is a more complex phenomenon than either spectral definition or SPL behavior for the characterization of an individual voice, because changes in vibrato are more intimately linked to expressive intentions. My remarks here apply only to a neutral vibrato of the kind most singers employ unconsciously.

We must consider two aspects of vibrato: vibrato rate and vibrato amplitude. Vibrato is a more or less regular change of pitch around a nominal center frequency; by vibrato rate I mean the frequency of this regular change, by amplitude its depth or excursion about this center pitch. Fig. 4a illustrates the vibrato of a sung Bb4.

You see that the note is attacked from below; the vibrato begins, increases to a maximum of about plus and minus 16 Hz, or more than a

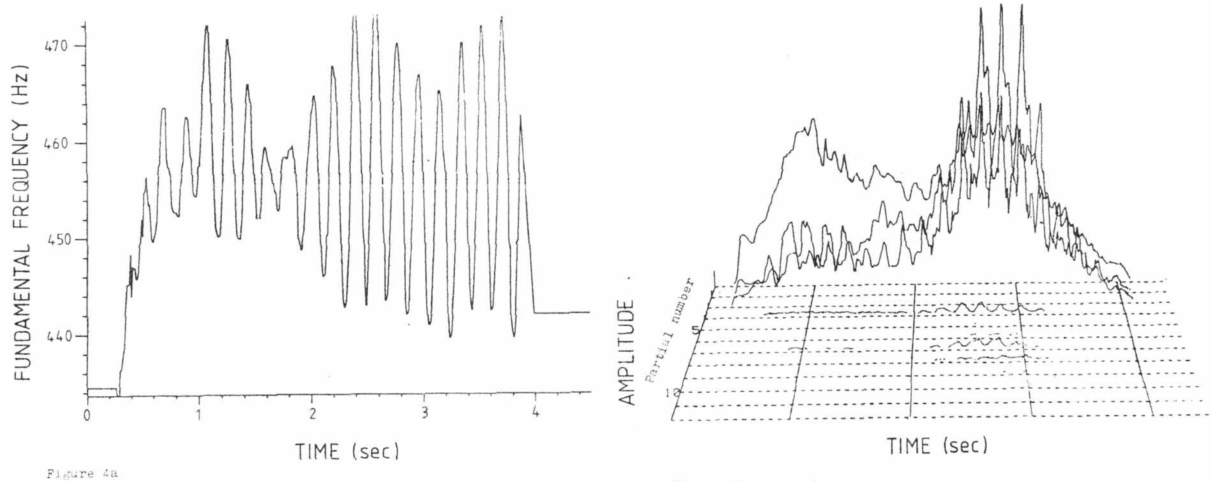

Figure 4a

Figure 4b

Fig. 4a and b.

quarter tone in each direction, then decreases and increases once more at the end of the note. The vibrato rate is rather slower at the beginning of the note, reaching an apparent target value (about 6 Hz) after approximately two seconds. Fig 4b shows the spectral evolution of the same note (low partials are at the back, higher ones in front, amplitude is in the vertical plane).

Here you see a strong initial attack in the first partial (maximum at about 0.9 seconds), then a decrease of amplitude followed by an increase with maximum at 2.5 seconds, and a final decrease of amplitude. You also see the entrance of the higher partials as the note gets louder, a very clear illustration of the way timbre changes with amplitude. (Note that the amplitude scale in this figure is linear and so gives a quite different picture of the relationship between lower and upper formant areas than do the other figures.)

By comparing Fig. 4a with Fig. 4b we can see to what extent vibrato amplitude is dependent on a note's fundamental amplitude. Only once do we see a discrepany between the two: at the end of the note the amplitude of the fundamental decreases rapidly while the vibrato amplitude increases. I have noted this increase in vibrato amplitude at the end of a note in most of the singers both male and female whom I have recorded. I imagine that the increased vibrato amplitude helps keep the soft sound interesting and relatively full in timbre.

In Fig. 4b we also see large fluctuations of amplitude in each of the partials having the same period as the vibrato variations in fundamental frequency. These fluctuations do not reflect changes in amplitude of the source signal synchronous with the changing fundamental frequency (there are none). Instead the changes are produced by the signal's passing through the narrow bandwidth resonances of the vocal tract: as a voice source partial changes frequency because of vibrato, it may move closer in frequency to the center of a resonance. As it does so, its amplitude increases, and as it moves away, its amplitude decreases again. Since the vocal tract's resonances are unlikely to be tuned to the partials of a note (although it seems that certain singers, particularly sopranos, make some effort to tune their formants to the nearest partials of long notes), different partials will reach their maxima at different times. In fact, this can be seen in Fig. 4b: compare partials 4 and 7 moving through two different resonances). The importance of the vibrato for a singer seems to me to lie in vibrato's ability to amplify the voice and to enrichen its timbre by ensuring that the partials near a formant actually bring that formant to greatest resonance.

In my arpeggio recordings I measured both vibrato rate and vibrato amplitude as a function of fundamental frequency. Fig. 5a and 5b summarize the results of these measurements.

The vibrato rate increases with higher pitch in all four voices up to a maximum, after which it decreases again. The pattern is much like that of amplitude evolution, although much more detailed studies than mine would be necessary to explore the relationship here. On the other hand, vibrato amplitude, which in Fig. 4a and 4b seemed clearly related to fundamental amplitude, acts in Fig. 5b much more like an independent factor, or rather like a factor more dependent on aesthetic choice than on the uncontrolled physical behavior of the voice. Three of the sopranos narrow the amplitude of their vibratos as they move up their ranges. One increases the amplitude of her vibrato with higher pitch, presumably because she prefers a deep vibrato on high notes and not because she cannot sing otherwise. The SOUND EXAMPLE 5 illustrates a synthesized voice singing the same phrase twice, with a different vibrato evolution each time.

4. Random variation of fundamental frequency

If we study Fig. 4a again carefully, we find that the fundamental frequency is by no means stable, but rather fluctuates in an irregular movement.

As far as we know, these fluctuations are truly random and reflect minute, uncontrolled variations in the tension of the muscles of the vocal folds. These small variations are essential to any convincing synthesis of the singing voice: without them even the most carefully placed formants and the most elegant vibrato sound machine-like; in their presence, the ear excuses small faults of timbre and vibrato. SOUND EXAMPLE 6 shows

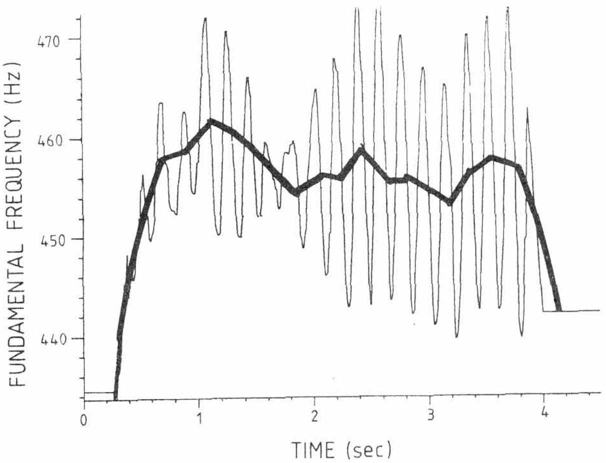

Fig. 6.

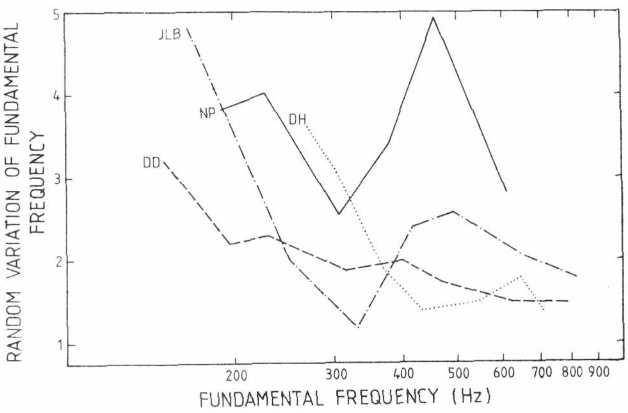

Fig. 7.

the importance of random variation of fundamental frequency: the two synthesized notes are identical except for the presence of random variation of fundamental frequency in the second note.

Fig. 7 shows the results of the measurement of random fundamental frequency variation as a function of pitch for our four sopranos.

In general the random variation decreases with ascending pitch. The largest variations are to be found at the lower end of the arpeggios, outside the actual soprano range, where control over pitch and timbre is naturally less precise because of the great relaxation of the vocal folds. As the voice moves up through the range, the tension of the vocal folds is increased, and the random variations of pitch become smaller.

Of the four aspects of the voice thus far discussed, random fundamental frequency variation seems to show the least difference in behavior between individual voices. Nevertheless, I think it important to model the general decrease in the excursion of the random variation with increasing pitch, and one can easily imagine designing a synthetic voice where the pattern of change of the random fundamental frequency variation would be an important part of the "personality" of the voice.

5. Attack patterns

In a general way, it seem that each singer has a repertoire of attack patterns, more or less extensive, depending on the attention he or she has given the matter. In my recordings, each of the singers had characteristic attacks. JLB, for example, usually overshoots the target pitch of a held note by between 8% at E4 and 4% at G#5 (the overshoot reaches 15% - 20% at the lower extreme of the range below C4). The length of the overshoot (that is the time for the beginning of the note until the target pitch is stable) is about 0.5 second at the lower end of the range and about 0.1 second at the upper end. Fig. 8a - 8f show the fundamental frequency patterns of the six notes in an E-major arpeggio from B3 to G#5. The vertical axis represent frequenccy in Hz, the horizontal time in seconds. Note particularly the overshoot at the beginning of each note and the irregular vibrato, compared with the vibrato in Fig. 4a.

Singer NB also attacks the notes in the lower part of her range from above (Fig. 9a-9d), but she takes much longer than soprano JLD to reach a stable pitch (from one to two seconds). She attack higher notes from below and reaches target pitch somewhat faster than for lower notes; her vibrato is much more regular than JLB's (cf Fig. 9e and 9f).

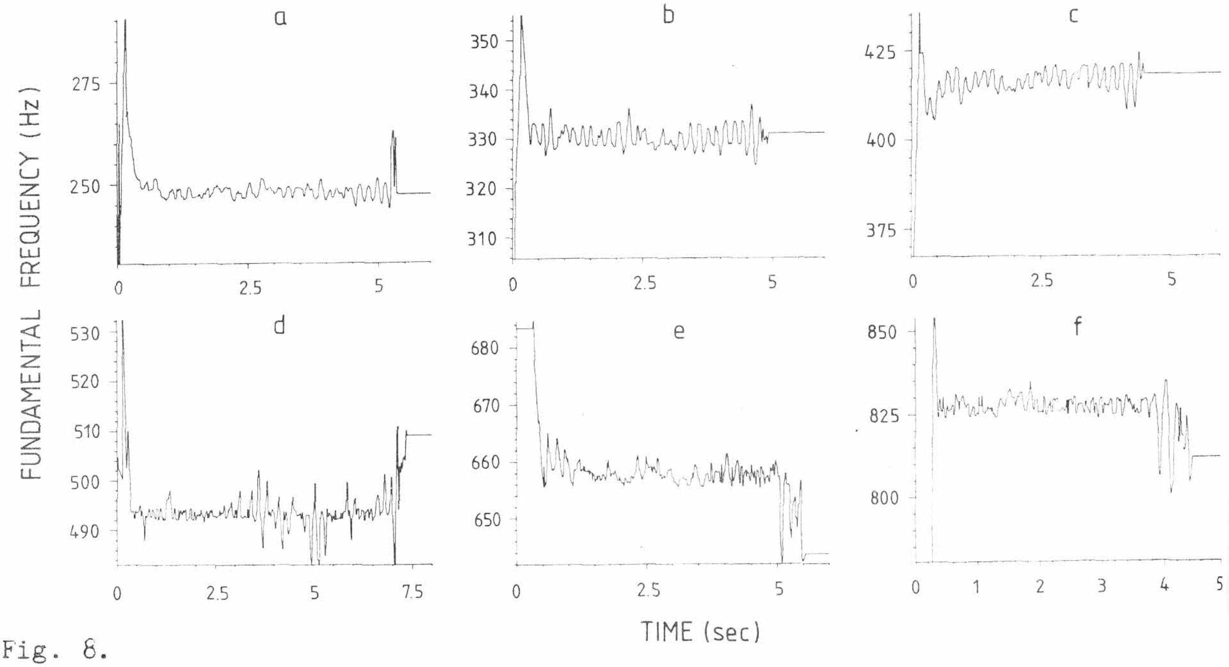

Fig. 8.

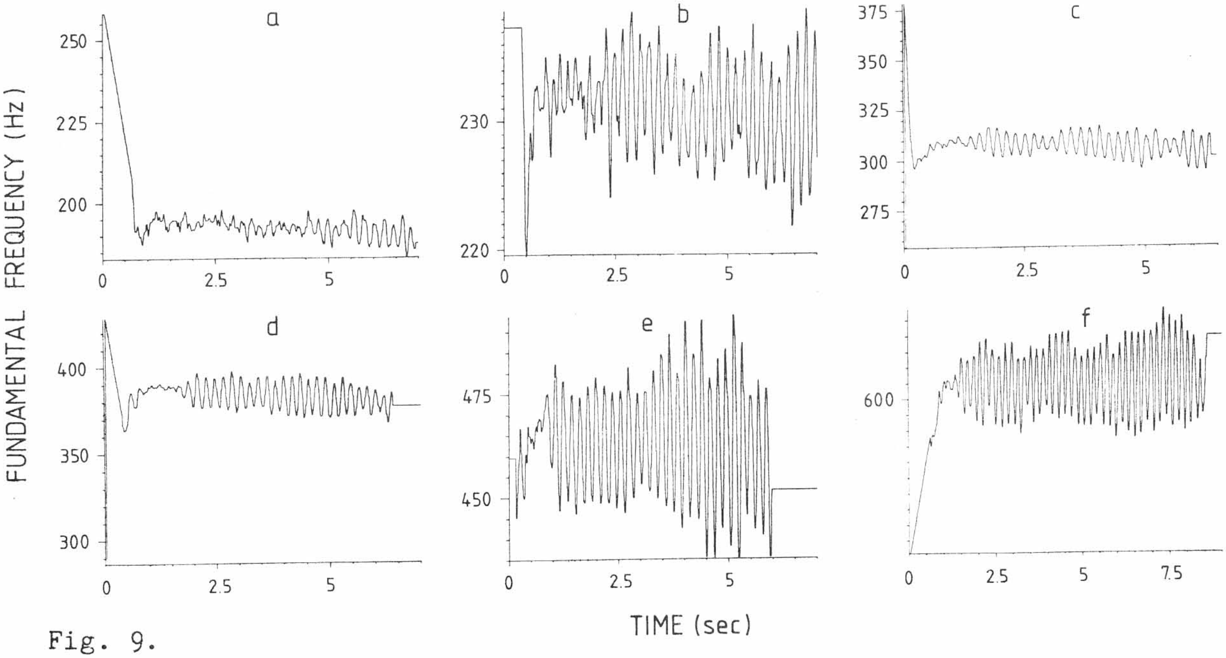

Fig. 9.

Clearly, a computer program offering dynamic control over pitch and amplitude can realize attack patterns like these without difficulty. But besides studying attack patterns, we must also consider the behavior of the global fundamental frequency of individual notes. Singer DD, for instance, always flattens fundamental frequency by a few Hz when she makes a crescendo.

6. Discussion

What is the interest of such analysis for electronic music? Of course, for me the most immediate interest has been in improving the synthesis of singing voices by computer. This has been the goal of all my work, and my research has been formulated to provide specific answers to specific problems of synthesis. But why did I choose to imitate singing voices at all?

My original interest in having the computer synthesize singing voices was compositional. Of course, straight imitation is not very exciting: there is no sense spending such time and effort on something singers will always do better than the computer. But I could imagine many compositional situations where I wanted an extension of the human singing voice, either in a direction no singer could go, or without conveying the sense of physical effort a human singer would expend to realize the effect. Here only synthetic voices would do; at the same time, the intrinsic quality, the complexity and the subtlety of the synthesized voices should be comparable to those of the natural voice; hence the interest in achieving a good imitation. But to imitate well necessitates understanding well what singers do when they sing; therefore studies like the present ones must be made.

This research has a more complicated relation to electronic music, less directly applicable than measurable results, to be understood rather in an illustrative sense. These studies give the composer a glimpse of how complex and how delicate the inner structure of sound is and show him interrelationships between physical dimensions of sound which may give guidance for constructing new and rich imaginary sonorous worlds.

Electronic music has always seemed to offer composers unlimited possibilities for making new sounds. The technical difficulty has always been to avoid using this richness in a merely anecdotal or decorative way, but rather to choose from the infinity of possibilities sounds varied and satisfying to the ear and yet unified at some deeper level. Many compser have experienced how synthetic sounds empirically chosen for a piece, each in itself interesting and complex, but without structural relation to the others, seem to loose much of their interest and complexity when combined. Here the ear seems very demanding, as if it were capable of performing analysis to determine the derivation and the inner constituency of sound. Johan Sundberg has spoken of how sensitive the ear is to the quality of "naturalness" in synthetic voices: the fact that most physical characteristics of the voice can vary only within strictly fixed boundaries; even the slightest transgression of the boundaries makes the voice sound "unnatural", while within these boundaries great variation is possible. Sundberg believes (if I understand him correctly) that this judgement is based on an analysis of how the sound is produced; if the analysis shows that the sound could not have been produced by natural means, we immediately loose a great deal of our discriminatory power for the sound.

I am certain that the same phenomenon occurs not only in imitations of nature but in electronic music in general: the perception's attempts to define the origin of a sound give insight into the sound's structure. A succession of sounds unrelated at this level may in many contexts be aesthetically less satisfying than a succession whose sounds have some kind of inner structural relationship. These studies of singing voices show some of the types os complexity the ear is apparently good at dealing with; thus they cast light on the organization of sound at a level deeper than mere surface interest, offering composers a conceptual model of the subtle nuances of sound's inner structure.

However, the research has had consequences more direct than simply to illustrate the complexity of the organization of sound. The discovery that most of the acoustical features important for defining an idividual voice are not unique values but dynamic factors dependent on the momentary context meant that we had to change our synthesis program so that it expected, in place of a single value for, say, vibrato rate, a rule allowing it to derive vibrato rate from the pitch and the intensity at the moment. This change in the program had an important consequence we had not originally foreseen: the ability to vary synthesis parameters dynamically and as a result of decisions taken by the program meant that the computer was no longer merely an instrument for sound production, but could be incorporated into the compositional process itself. This is certainly not the first time composers have used the computer to help them write music, but whereas the computer generally does little more than produce a score to be played either by traditional instruments or by a separate synthesis program, in our program one can pass on as much control to the computer as one wishes, from the creation of a score to the adjustment of the smallest detail of the sound synthesis itself.

This research into singing has, we believe, led to a significant advance in both the theory and the practice of sound synthesis. Our model, in contrast to most sound synthesis programs, is a physical model, that of the vocal tract. Certain acoustical features of the voice have been found to be more closely interrelated than others. These interrelationships have defined the stucture of the program. We believe strongly that these relationships will keep their validity even when the program is used to produce sounds bearing no apparent similarity to vocal sounds.

Let us conclude by listening to SOUND EXAMPLE 7, a short section of a piece for tape alone, *Winter(1980)*. Here the synthesis program produced both singing voices and sounds which do not sound vocal. The musical organization of such heterogeneous material was satisfyingly strict; since both vocal and non-vocal sounds had the same general structure, I could subject both, despite their superficial differences, to the same operations. Thus it was the synthesis program which suggested many of the aspects on which to base the piece's musical development, namely those acoustical features lining apparently very different materials at a deep structural level. But it was also the synthesis program which organized much of this diverse material. Here musical imagination and acoustical analysis moved forward together, each inspiring the other, each enrichened by the other's experience.

REFERENCES

FANT, G. (1970): *The Acoustic Theory of Speech Production*, Mouton, The Hague

FANT, G. (1973): *Speech Sounds and Features*, The Massachusetts Institute of Technology Press, Cambridge

MATHEWS, M. V. (1970): *The Technology of Computer Music*, The Massachusetts Institute of Technology Press, Cambridge

RODET, X. & BENNETT, G. (1980): "Synthèse de la voix chantée par ordinateur", Confrences des Journées d'Etudes 1980, Festival International du Son, 73-91

SUNDBERG, J. (1978): "Synthesis of singing", Swedish Journ. Musicology *60:1*, 107-112

SUNDBERG, J. (1979): "Perception of singing", Speech Transmission Laboratory Quarterly Progress and Status Report 1/1979, 1-48

SUNDBERG. J. (1981) Personal communication