STM 1978:1

Synthesis of singing

By Johan Sundberg

Synthesis of singing*

By Johan Sundberg

Introduction

Recently, a terminal voice analogue has been constructed in the Speech Transmission Laboratory, Department of Speech Communication, Royal Institute of Technology (KTH), Stockholm (Larsson 1977). The equipment is designed to meet the particular demands encountered in synthesizing singing. The purpose of the present paper is to show that such an analogue may be useful in attempts to elucidate psychoacoustical aspects of singing. After a presentation of the equipment, the timbral effects of a formant technique observed in a soprano's singing of high-pitched tones will be demonstrated. Then the perceptual effects are illustrated of various acoustic data observed in a male professional singer's performance of a song.

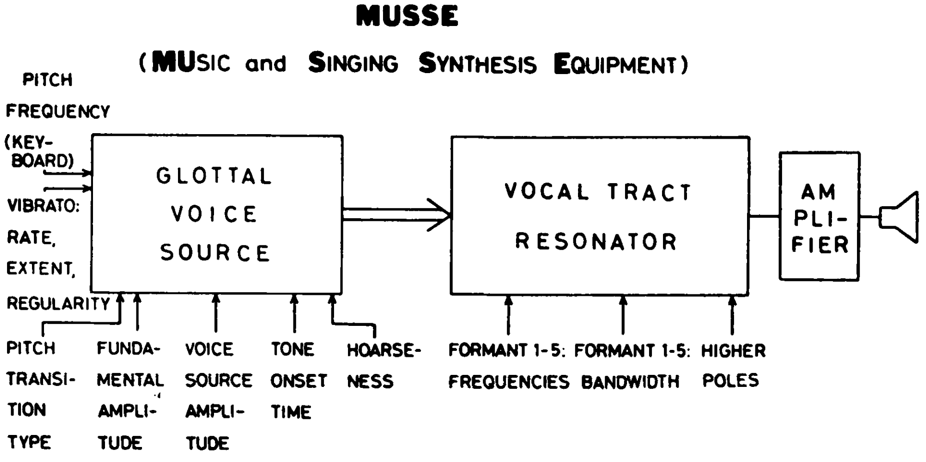

Fig. 1.

The MUSSE synthesizer

The analogue is called Music and Singing Synthesis Equipment, or in short, MUSSE. As is shown in Fig. 1, MUSSE consists of two major blocks. One is an analogue to the vibrating vocal folds, and the other models the vocal tract resonator. Thus, as yet, only voiced sounds can be synthesized.

The vocal tract parameters, i.e. the formant frequencies, substantially affect

* Paper given at the IRCAM symposium on the Psychoacoustics of music, Paris 11—13 July, 1977.

the timbral properties of the tone. In *sound example 1,*[1] the formant frequencies are changed one by one, according to the following scheme:

(a)	first	formant frequency	shifts from	880	to	560	Hz
(b)	second	—„—	—„—	1770	to	970	Hz
(c)	third	—„—	—„—	2050	to	2350	Hz
(d)	fourth	—„—	—„—	3860	to	2680	Hz
(e)	fifth	—„—	—„—	4910	to	2910	Hz

The result is a reasonably acceptable [*a*] vowel (as in the word wather). However, its naturalness suffers from the static characteristics of the voice source.

In *sound example 2,* the voice source parameters are changed in the following order:

(a) the vibrato rate shifts from 4 to 12 to 6 undulations per sec,
(b) the vibrato extent shifts from ± 0% to ± 12% to ± 3% of the mean fundamental frequency,
(c) the vibrato rate regularity, i.e. the standard deviation of the undulation cycle time, shifts from 0 to 20 to 2.5%.

The way in which the pitch changes from one value to another is also decisive for the naturalness of the synthesis. This can be heard in *sound example 3,* where the pitch changes a) in steps, b) rather slowly in a glissando-like manner, and c) as in normal trained singers: the central 75% of the pitch change is completed in 70 msec (cf. Sundberg 1975).

The relative amplitude of the voice source fundamental can be changed so that its amplitude varies between +9 dB and +2 dB relative to the amplitude of the second voice source partial. These cases can be listened to in *sound example 4.*

In *sound example 5,* the tone onset time is first set to 25 msec and then to 400 msec. Note that the onset of the vibrato is always delayed about 600 msec as soon as the tone is preceded by a rest.

Sound example 6 illustrates the effect of adding noise pulses in synchrony with the glottal wave as suggested by Rothenberg (1974).

After this presentation of the MUSSE parameters, some examples of a more complex synthesis will be presented. During the performance of even simple tone sequences a singer may change most of the parameters which are modelled in MUSSE. Hence, synthesizing such sequences requires a computer control of MUSSE. The program used was developed by Carlson and Granström with the object of converting a written text to speech (cf. Carlson and Granström 1975). That purpose is similar to the present one, converting a string of note signs into tone sequences. In both cases the conversion must be made by applying a set of explicitly formulated rules, from which the computer calculates the sound. The output sound then reveals to what extent the rules agree with the rules that consciously or unconsciously are applied by a person who reads a text aloud or, in our case, sings. Here, then, the synthesizer helps us to find out what rules relate the notation to sound in a musical performance.

[1] The sound examples are to be found on the record accompanying this issue of the journal.

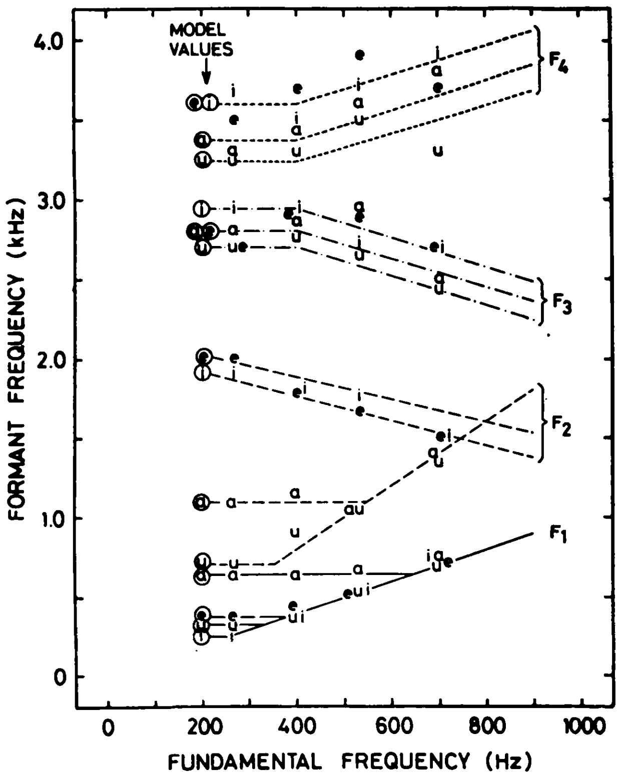

Fig. 2.

Female high-pitched singing

In high-pitched singing, female singers tend to increase their jaw opening with rising pitch. This is contrary to normal speech, where the jaw opening depends mainly on the vowel. According to acoustic theory, an increase of the jaw opening tends to raise the frequency of the first formant. Acoustic measurements on one professional soprano singer have revealed that the pitch-dependent jaw opening seems to serve a very special purpose: to avoid that the first formant is lower in frequencly than the fundamental. The strategy seems to be that first formant is tuned to a frequency close to the pitch frequency, if otherwise the pitch would exceed the first formant in frequency (cf. Sundberg 1975). Thereby, all other formant frequencies are changed too, as is illustrated in Fig. 2.

Fig. 2 shows the frequencies of the four lowest formants as functions of the fundamental frequency. The non-circled symbols show the values obtained from the soprano singer, and the circled symbols and the straight lines show approximations of the same data. These approximations were tested on MUSSE; they were expressed as rules relating the formant frequencies to the vowel type and the fundamental frequency. The first half of *sound example* 7 presents the sound

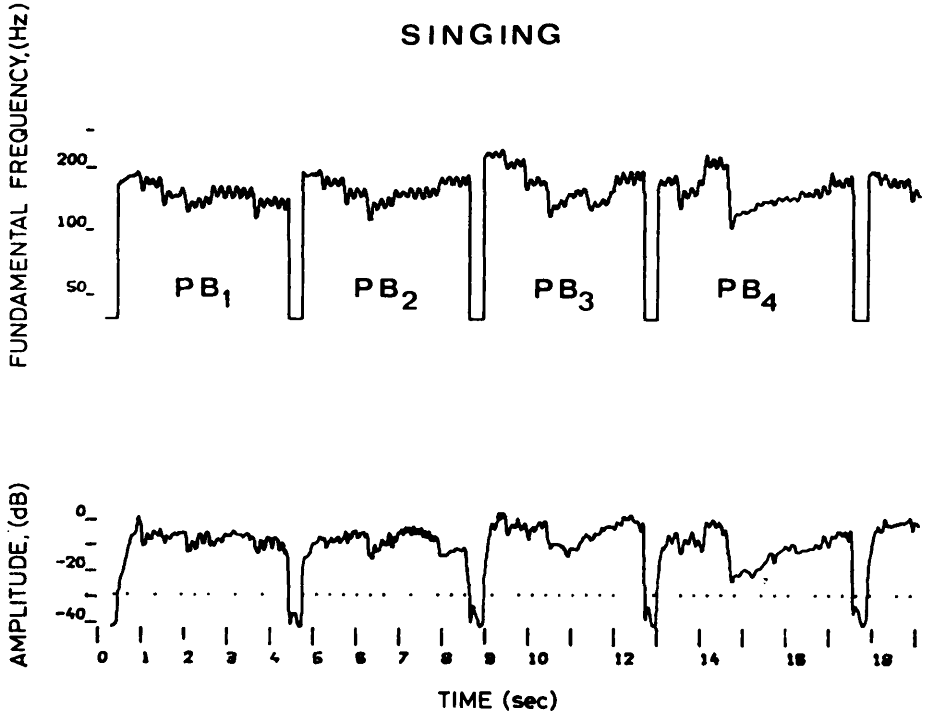

Fig. 3.

that would result if the soprano did *not* change her formant frequencies with the pitch. The sound is faint and highly unnatural. The second half of the same example gives the sound resulting fram the straight line approximations shown in the figure. This is certainly a better approximation of what is typical of professional female singing. Note also that the loudness increased substantially even though the voice source amplitude remained the same in both sets of vowels. This means that the pitch-dependent jaw opening is a way to vocal economy in female high-pitched singing. Thus, we can conclude that the straight line approximations seem to possess a certain degree of generality, even though they were based on data obtained from one single subject only.

"Musical" performance

The last set of sound examples is related to the acoustic bases of musicality in a performance of a song: what are the rules that a singer consciously or unconsciously applies, when he or she converts a set of note signs into *music,* as opposed to triads and scales? We can make guesses about this on the basis of acoustic data from a performance. Moreover, by means of synthesis we can explore the perceptual implications of our guesses, if we formulate them as rules controlling the synthesis.

A recording was made of a simple song (Panofka: *Vokalise,* G flat major, Op. 90:2) in an anechoic chamber. The singer is one of the most famous baritones in Sweden. In order to provide reasonably realistic experimental conditions, he heard not only the piano accompaniment but also his own voice in earphones. His performance was analyzed with respect to fundamental frequency and overall amplitude. Fig. 3 shows an example of the analysis used. It shows fundamental frequency and overall amplitude as functions of time. The graph was obtained by means of a computer program developed by Askenfelt (cf. Askenfelt et al. 1977). PB denotes *Pair of bars,* which is an important structural unit in this composition.

In the first example (*sound example 7a*), all parameters except the pitch are constant. In the following examples this synthesis is made more complex by introducing, step by step, major characteristics that were found in the singer's rendering of this song.

The singer typically approaches a following lower pitch from below, i.e. with a sort of undershoot. When this feature is introduced in the synthesis the result is as in *sound example 7b.* — The amplitude was observed to increase with pitch to a certain extent. If we add this feature to the synthesis we obtain *sound example 7c.* — Most notes were produced with a small crescendo—decrescendo of their own. If we add such a rule to the synthesis we obtain *sound example 7d.* — Towards the end of each pair of bars, the singer replaced the one-note long crescendo—decrescendo-gesture by a longer gesture of the same kind. In most pairs of bars a sudden decrease in amplitude occurred on the fourth note and the amplitude was then increased to a maximum on the sixth quarter note beat in the pair of bars. Adding this rule to the synthesis brings it to the status of *sound example 7e.* — There is one evident exception to this last-mentioned rule. It occurs at the end of the fourth pair of bars, where there is no decrescendo. The effect of adding this and a piano accompaniment to the synthesis can be heard in *sound example 7f.* It seems that each step here has brought the synthesis closer to a performance that is acceptable from a musical point of view, even though much more work is needed before the synthesis can compete with that of a real professional singer.

Conclusions

A classical problem in acoustics and hence also in singing research is that there is no simple relationship between acoustic data and what we perceive. Here, an attempt has been made to show that synthesis of singing is a promising avenue if we want to explore such relationships, i.e. to understand perceptually the acoustic data, and to find out what it is that makes singing sound like singing. The first example demonstrated the significance of formant frequencies, vibrato, and other voice parameters to voice timbre. The second example showed that singing synthesis may help us to determine the generality of observations made on one single subject only. The third example allowed us to experience how various acoustic characteristics observed in a musical performance of a song contribute to the musicality of the performance.

It is a general experience in work with synthesis of well-known sounds that the results generate new questions. Mostly these questions are quite explicit. Hence they can easily be answered. In this way synthesis is a highly rewarding procedure in scientific research. Let us just mention one example. In the soprano synthesis, there is a slightly screaming quality to the top pitch. Does this signify that it is impossible for a soprano to raise her formant frequencies this high in a "relaxed" manner? And how should we choose the formant frequencies in order to produce a perfect copy of a specific soprano singer? More measurements and continued synthesis work can answer these questions. Thus, it seems that MUSSE is a powerful tool in singing research.

Acknowledgment: The author is indebted to Rolf Carlson and Björn Granström for their assistance in the programming work. The work was supported by the Bank of Sweden Tercentenary Foundation (RJ).

References

Askenfelt, A., Gauffin, J., Kitzing, P., and Sundberg, J. (1977): Electroglottograph and contact microphone for measuring vocal pitch. A comparison. (Speech Transmission Laboratory, Quarterly progress and status report 1977/4, pp. 13—21, KTH, Stockholm.)

Carlson, R. & Granström, B. (1975): A phonetically oriented programming language for rule description of speech. (Speech communication, ed. G. Fant, Almqvist & Wiksell, Stockholm, Vol. 2, 1975, pp. 245—253.)

Larsson, B. (1977): Music and singing synthesis equipment (MUSSE). (Speech Transmission Laboratory, Quarterly progress and status report 1977/1, pp. 38—40, KTH, Stockholm.)

Rothenberg, M. (1974): Glottal noise during speech. (Speech Transmission Laboratory, Quarterly progress and status report 1974/2—3, pp. 1—10, KTH, Stockholm.)

Sundberg, J. (1975): Formant technique in a professional female singer. (Acustica 32, 1975, pp. 89—96.)