

Synthesising Singing

Johan Sundberg

Department of Speech Music Hearing, KTH Stockholm

Abstract - This is a review of some work carried out over the last decades at the Speech Music Hearing Department, KTH, where the analysis-by-synthesis strategy was applied to singing. The origin of the work was a hardware synthesis machine combined with a control program, which was a modified version of a text-to-speech conversion system. Two applications are described, one concerning vocal loudness variation, the other concerning coloratura singing. In these applications the synthesis work paved the way for investigations of specific aspects of the singing voice. Also, limitations and advantages of singing synthesis are discussed.

I. INTRODUCTION

Describing sound is difficult. Even though a number of words are commonly used in such descriptions, it is generally extremely difficult or even impossible to imagine the sound described. Part of the problem would be that many of the adjectives are relative, so a “high tone” simply means that the note is higher than some other tone, the pitch of which is typically not specified. Other words are imprecise in that they refer to properties of physical rather than of sonic objects, such as “bright”, “harsh” or “raw”, or to human moods, such as “aggressive”.

These difficulties represent a major problem in music science, since music is art built by sound. The problem is particularly embarrassing in music acoustics, where one of the important tasks is to describe and explain how instruments sound and why they sound as they do.

When my interest in music acoustics started, it was common practise in much organology to describe, for example, organ timbre in terms of “moon shine”, “rattling birch leaves” etc. This demonstrated a need for more scientific methods.

Acoustic analysis appeared as an attractive alternative, as it yields quantitative and reproducible data. Moreover, it seemed fair to assume that other scientists would be able to understand and interpret such descriptions in perceptual terms.

On the other hand, acoustic analyses offer an overwhelming amount of data, and some of them totally lack perceptual relevance. In fact, as a sound analyser the human ear is rather poor. For example, most partials in a spectrum are inaudible, because strong spectrum partials mask other partials. Furthermore, the ear is insensitive to phase, so two waveforms that look entirely different may still sound exactly the same (Kakusho & al., 1968). In addition the limits of the audible frequency range are much narrower than those of most acoustic analysis devices.

As a consequence, a description of a musical sound is likely to contain large amounts of perceptually irrelevant information.

Analysis by synthesis represents an attractive solution to such problems. It is a classical research method in speech research. Jean Claude Risset was one of the first to apply this method to analysis of musical sounds (Risset, 1965, Risset & Mathews, 1969). His analysis revealed a simple principle which explained how the amplitude of the spectrum partials of trumpet tones varied when loudness was varied. He then implemented this principle in quantitative terms into a computer algorithm which he used to synthesise trumpet sounds. The resulting tones were practically impossible to distinguish from real trumpet tones. In this way, Risset had demonstrated that his algorithm represented a perceptually exhaustive description of trumpet sounds.

There are unique advantages with analysis-by-synthesis. One is that you can find out what acoustic properties are the salient ones. Another advantage is that you do not need a terminology for describing the timbral properties of the instrument. It is enough that you know how the instrument sounds so that you can compare it with the sound of the synthesis. A third advantage is that working with sound synthesis tends to draw your attention to details that may be quite important, even though they mostly pass unnoticed. Listening to the synthesis helps to direct your attention to such characteristics, and then, it is possible to define a term for them. The aim of this presentation is to review some examples of this.

II. THE MUSSE SINGING SYNTHESISER

My attempts to use analysis by synthesis in research on musical sounds started in the 1970s. Speech research had convincingly demonstrated that synthesis is a powerful tool in scientific research and Gunnar Fant and his department at KTH, at that time called the Speech Transmission Laboratory, had reached a leading international position in the acoustics of voice production as well as in the area of speech synthesis. As I had completed my doctoral dissertation at that department, singing synthesis was a natural thing to try.

The start was the realisation, in terms of a thesis work, of an idea of my department colleague Jan Gauffin to construct a hardware singing synthesiser. It was designed as a musical cousin to Gunnar Fant's classical speech synthesiser OVE (“Orator Vox Electrica”). One of Gauffin's ideas was that formant frequencies and other synthesis variables should be continuously variable rather than variable in small but

discrete steps. The idea was realised in terms of photo resistors controlled by the brightness of an electret light. The result was called the KTH Music and Singing Synthesis Equipment, or MUSSE (Larsson, 1977).

MUSSE was played from a keyboard and was provided with a number of knobs and switches that controlled properties relevant to singing. Apart from five formant frequencies and bandwidths, vibrato rate and extent, pitch-synchronous glottal noise, random variation of fundamental frequency F0, and rate of F0 change between notes could be varied by knobs. Formant amplitudes were controlled by algorithms, but also by the formant bandwidths, just as in the human voice. In addition, some variables could also be controlled by a joy-stick.

Some years later possibilities were created to control MUSSE also by digital signals (Malmgren, 1978). A modified text-to-speech conversion system, RULSYS, developed within the department (Carlsson & Granström, 1975) was used to control MUSSE via this interface. This allowed the conversion of input music files into performances of *vocalises*, i.e., songs sung on sustained vowels rather than with lyrics. Such songs are frequently used in teaching singing. The input file contained information on vowel, pitch and tone duration, and the modified RULSYS program converted this information to formant frequencies, amplitude, timing and vibrato parameters.

The experiences from the MUSSE synthesis were quite important. The synthesis demonstrated that the MUSSE synthesizer could produce synthesis of excellent vocal quality, but also that the result was a disaster from a musical point of view. The lack of evidence in the performance of an urge to communicate and to express something that the (imagined) singer felt as exceedingly important or fascinating became painfully evident.

These experiences was a striking demonstration of the relevance of musical expression in music performance, a relevance that, remarkably enough, is often totally neglected in today's music culture where computers are commonly used to play dead-pan versions of music.

Attempts were made in collaboration with Rolf Carlson and Björn Granström in the speech group of the department to cure this deficiency of MUSSE performances. The strategy was to take advantage of the context dependent rule tool that they had incorporated into their RULSYS program for text-to-speech conversion synthesis. By implementing context dependent accent, phrasing and marcato rules, the lifeless character of MUSSE's dead-pan performance of a Vocalise by Panofka could be efficiently reduced, particularly when a live piano accompaniment was added.

Possibilities to synthesise consonants were established some years later in terms of a thesis work (Ponteus, 1979). The first synthesis concerned the solmisation syllables and was carried out mainly by Jan Zera, a Polish guest researcher (Zera et al., 1984).

A. Crescendo and diminuendo

A particularly striking experience from the MUSSE synthesis was its failure to create realistic dynamic variation. Originally crescendo and diminuendo were modelled by variation of the amplitude of the voice source waveform. This produced a peculiar effect. Crescendos and diminuendos sounded as if the singer varied the microphone distance rather than vocal loudness. The trick to achieve a realistic sounding synthesis of an increase of vocal loudness was to decrease the spectrum slope of the voice source. This effect was implemented in the MUSSE system in terms of a physiologic volume control unit, which was controlled by the same signal as that used for controlling the voice source amplitude. The background was that the main correlate of an increase of vocal loudness is a reduction of the overall spectrum slope, as illustrated in Figure 1.

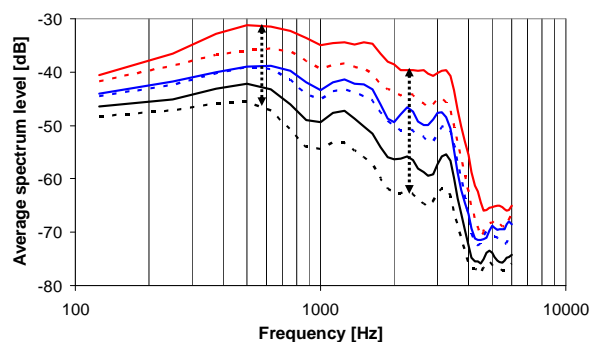


Figure 1. Long term average spectra of a male subject reading the same text at different degrees of vocal loudness. The level difference near 500 Hz is 14 dB and that near 2500 Hz is 22 dB, thus illustrating the effect of loudness variation on the spectrum slope of the voice source. From Nordenberg & Sundberg, (2004)

B. Coloratura

Coloratura is a sequence of short notes, often 16th notes that are sung on one single vowel. It was abundantly used in arias during the Baroque era, and remained a prominent item in opera music up to the end of the 19th century. Analyses of coloratura sequences revealed that they are typically performed as illustrated in Figure 2. F0 makes a turn around each target frequency. The tempo of about six sixteenth notes per second implies that the duration of each rise-fall cycle is about 170 ms, which is quite similar to the cycle time of the vibrato undulations in singing. Indeed, the F0 pattern appears as a vibrato combined with a glissando. However, at turning points in the melodic line, the frequency made a large overshoot for the top note.

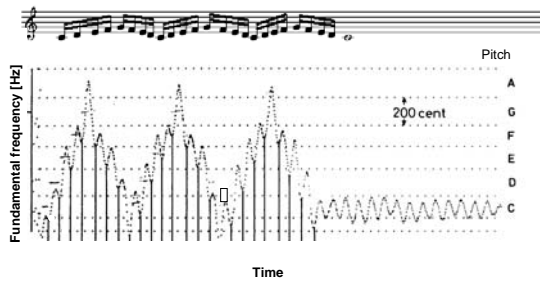


Figure 2. Fundamental frequency pattern in a professional singer's performance of the coloratura sequence shown at the top. The dotted lines show the boundaries midway between the scale tones shown to the right.

Implementing this simple principle in the MUSSE synthesis showed that this F0 pattern indeed produced a realistic-sounding coloratura (Sundberg, 1981). Even the overshoot at the peak tone at melodic turning points was important. Without it, the top note sounded flat.

This experience raised several questions. How can such F0 patterns be produced? Could they result from superimposing a vibrato on a glissando? And why do we hear such continuously varying F0 patterns as a sequence of discrete pitches?

A typical example of a physiological correlate of coloratura sequences can be seen in Figure 3, showing how the oesophageal pressure, captured by means of a pressure transducer, was varied when a singer performed the coloratura sequence shown at the top in the same figure. The changes in this pressure reflect the changes of the subglottal pressure, i.e., the air pressure in the respiratory system under the glottis, which drives the vocal fold vibrations. It can be seen that each tone in the sequence of sixteenth notes is produced with a pressure pulse and that the pressure pulses are synchronised with the F0 undulations. Thus, in the F0 curve each note corresponds to a small rise and fall. It would take quite special skills to produce such carefully synchronised patterns of subglottal pressure variation and F0 variation.

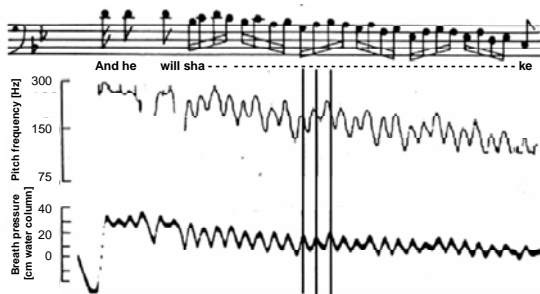


Figure 3. Fundamental frequency and oesophageal pressure observed when a professional baritone singer performed the coloratura sequence shown at the top.

The pressure variations are not likely to result from a modulation of glottal adduction. A weak glottal adduction should result in a voice source dominated by the fundamental and with very weak high spectrum partials. Mostly the amplitudes of the high overtones do not vary in coloratura sequences. Rather the

pressure variations would be caused by pulsating contractions of respiratory muscles.

An increase in subglottal pressure causes F0 to increase by a few Hz per cm H₂O (see e.g. Titze, 1989). This implies that a subglottal pressure that pulsates with an amplitude of, say, ± 5 or ± 7 cm H₂O will cause a F0 pulsation of about ± 15 or ± 20 Hz. In the example shown in the figure, F0 varies between 120 and 250 Hz, approximately. In this frequency range a semitone corresponds to 8 to 16 Hz. This suggests that, indeed, the F0 undulation in the coloratura passage was perhaps caused by the pulsation of the subglottal pressure. If this is correct, the singer would generate a coloratura passage by combining a glissando, produced with glottal pitch raising muscles, with a pulsation of the subglottal pressure, produced with the respiratory apparatus.

How is it possible that the F0 pattern of coloratura is perceived as a sequence of discrete pitches? The answer seems related to the phenomenon that we perceive a well-defined pitch of vibrato tones, provided the vibrato rate and extent are kept within certain limits. Thus, if the modulation cycle of the vibrato is near 170 ms, i.e., if the vibrato rate is near 6 Hz, a clear pitch is perceived that corresponds to the average frequency of the vibrato tone. This suggests that the ear's pitch perception is based on low-pass filtering of the F0 signal at about 6 Hz. If such an averaging is applied to the frequency pattern underlying a coloratura passage, an F0 pattern of the type shown in Figure 4a should produce the perception of a sequence of discrete pitch steps, as illustrated in Figure 4b.

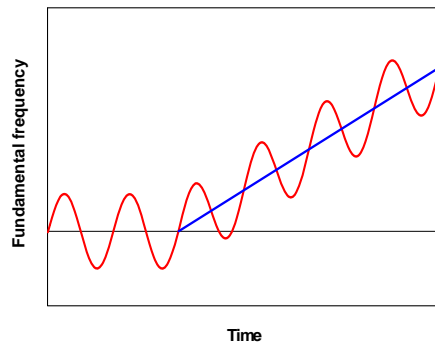


Figure 4a: Illustration of the result of superimposing a sine wave signal on a ramp function.

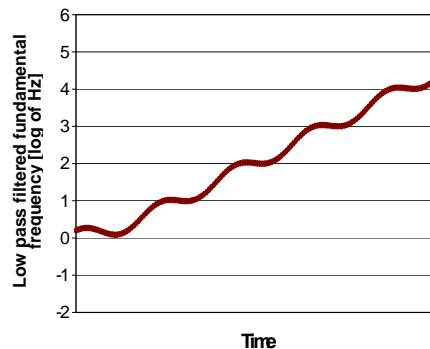


Figure 4b (right graph): The same signal after low pass filtering at 6 Hz.

III. DISCUSSION AND CONCLUDING REMARKS

Experiences from the analysis-by-synthesis work with singing has had a great impact on research. The hardware MUSSE was replaced by a software version during the 90s. This implementation was realised by Sten Ternström. Context-dependent rules for music performance were developed over a period of more than 20 years under the musical guidance of a leading Swedish performance teachers, professor Lars Frydén. Anders Friberg beautifully developed the results and implemented them into the Director Musices program (Friberg, 1995). The Director Musices program allows you to modify, even in real time, the performance of an input music file by a set of performance rules. Director Musices is now available as freeware on the internet (<http://www.speech.kth.se/music/performance/download/>).

The principle of synthesising dynamic variation by varying the spectrum slope of the voice source has been documented in trained and untrained voices (Nordenberg & Sundberg 2004; Sjölander & Sundberg, 2004). It has been explained and modelled by Fant (Fant et al., 1985). Nevertheless, the substantial effects of vocal loudness variation on voice characteristics is still commonly neglected in voice measurements.

The yield offered by voice synthesis is significantly affected by the construction of the synthesiser. Sound recording systems represent one extreme. They can synthesise the original sound very accurately, but it is difficult to draw scientific conclusions from working with this type of synthesis. The opposite extreme is represented by physiologic models. They are composed by components modelling the actual parts of the voice organ. Thus, the vocal folds are represented by computerised replicas of the laryngeal tissues, and the vocal tract is a filter, the resonance characteristics of which are controlled by equations corresponding to the positions of the articulators, jaw, lips, tongue, velum and larynx.

This appears to suggest that the best models are the most realistic ones. On the other hand, such models tend to be extremely difficult to control, since they contain so many components. They require precise specifications of the behaviour of every single detail of the system, and in such situations it is easy to compensate one mistake by introducing another one. The sound recording system, is of course extremely easy to control, but it rarely teaches you anything about how the singer produced the sound.

Given this dilemma research needs to make a smart choice. Thereby a principle launched by the logician William of Occam, born around 1280 and dead around 1350, still applies. The principle, commonly referred to as *Occam's razor*, states that the explanation of any phenomenon should make as few assumptions as possible, "shaving off" those that make no difference in the observable predictions of the explanatory hypothesis or theory. The Latin name of the principle is *lex parsimoniae*, or the law of parsimony, meaning that the simplest explanation of observed phenomena should be preferred. Occam's razor principle is widely honoured in scientific

research, but not always in voice research, where the urge for realism and completeness is sometimes ranked higher than preference for simplicity.

However, it is sometimes not obvious how the simplicity criterion should be applied. For example, we may compare two models for voice synthesis, one that works with the frequencies of the formants, like the MUSSE synthesiser, and one that works with the movements of the articulators, as the APEX model freeware available at:

<http://www.speech.kth.se/~pjohan/currentprojects.html>

Both models are quite complex, but explanations of formant frequency changes in terms of shifts in e.g., tongue shape and jaw opening are certainly much easier to understand than explanations describing nothing but how the formant frequency change in various contexts.

Summarising, analysis-by-synthesis as applied to singing has proved to be a quite rewarding research tool. The examples presented here have demonstrated its potentials of shedding light on the acoustic correlates of dynamic variation, on pitch perception and on music communication in general. There are reasons for an optimistic view of the results of applying the analysis-by-synthesis strategy to the singing voice in the future.

IV. ACKNOWLEDGEMENTS

Part of the material presented in this article was recently published elsewhere (Sundberg, 2006).

REFERENCES

- [1] Carlson, R., Granström, B. (1975). A phonetically oriented programming language for rule description of speech. In G. Fant (ed.), *Speech Communication* (pp. 245-253), Stockholm: Almqvist & Wicksell, vol 2. 38-40.
- [2] Fant G, Liljencrants J, Lin QG (1985) A four-parameter model of glottal flow, *Dept. for Speech, Music and Hearing Quarterly Progress and Status Report* 26:4, 1-13. (http://www.speech.kth.se/prod/publications/files/qpsr/1985/1985_26_4_001-013.pdf)
- [3] Friberg A (1995) *A Quantitative Rule System for Musical Expression*, doctoral dissertation (Music Acoustics), KTH 1995.
- [4] Kakusho O, Kato K, Kobayashi T (1968) Just discriminable change and matching range of acoustic parameters of vowels, *Acustica* 20, 46-54.
- [5] Larsson, B. (1977). Music and singing synthesis equipment (MUSSE). *Speech Transmission Laboratory Quarterly Progress and Status Report*, 1/1977, 38-40.
- [6] Malmgren, J. (1978). *MUSSE Interface Unit MIU*. Thesis work, Department of Speech Music Hearing, KTH.
- [7] Nordenberg M, Sundberg J (2004) Effect on LTAS on vocal loudness variation. *Logopedics Phoniatrics Vocology* 29, 183-191.
- [8] Ponteus, J. (1979) Mimmi, an equipment for consonant synthesis intended as a complement for MUSSE (in Swedish), Thesis, Department of Speech Music Hearing, KTH.
- [9] Risset JC (1965), Computer study of trumpet tones, *J. Acoust. Soc. Am.* 38, 912.
- [10] Risset JC, Mathews MV (1969): Analysis of Musical Instrument Tones, *Physics Today*, 22, Feb. 23-30.
- [11] Sjölander P & Sundberg J (2004) Spectrum effects of subglottal pressure variation in professional baritone singers, *J Acoust Soc Amer* 115, 1270-1273

- [12] Sundberg J (1981) Synthesis of singing, in *Musica e Tecnologia: Industria e Cultura per lo Sviluppo del Mezzogiorno*, Proceedings of a symposium in Venice, Venedig: Unicopli, 145-162.
- [13] Sundberg J (2006) The KTH synthesis of singing, *Advances in Cognitive Psychology, Special Issue on Music Performance*, 2-3, 131-143.
- [15] Titze I (1989) On the relation between subglottal pressure and fundamental frequency in phonation_ *J. Acoust. Soc. Am.* 85, 901-906.
- [16] Zera J, Gauffin J, and Sundberg J (1984) Synthesis of Selected VCV-Syllables in Singing, *Proc. International Computer Music Conference*, IRCAM, Paris, 83-86.