2006 • volume 2 • no 2-3 • 131-143

Advances in Cognitive Psychology

The KTH Synthesis of Singing

Johan Sundberg

Department of Speech, Music Hearing, School of Computer Science and Communication, KTH, Stockholm

Received 27.02.2006
Accepted 07.08.2006

Keywords
singing voice, formants, voice source, music performance, analysis-by-synthesis

ABSTRACT

This is an overview of the work with synthesizing singing that has been carried out at the Speech Music Hearing Department, KTH since 1977. The origin of the work, a hardware synthesis machine, is described and some aspects of the control program, a modified version of a text-to-speech conversion system are reviewed. Three applications are described in which the synthesis system has paved the way for investigations of specific aspects of the singing voice. One concerns the perceptual relevance of the center frequency of the singer's formant, one deals with characteristics of an ugly voice, and one regards intonation. The article is accompanied by 18 sound examples, several of which were not published before. Finally, limitations and advantages of singing synthesis are discussed.

INTRODUCTION

Describing sound is difficult. Even though there are a number of words available that describe sound, it is generally impossible or at least extremely difficult to hear internally the sound described. Part of the problem would be that many of the adjectives are relative, so a "high note" simply means that the note is higher than some other tone, the pitch of which is typically not specified. Other words are not very precise, such as "harsh", "raw", "hoarse". It is simply quite difficult to realise how a sound described in such terms really sounds.

This difficulty of describing sounds is a major problem in music acoustics, since one of its main research areas is the sound of musical instruments. Hence, the ultimate task is to describe and explain how they 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, or rattling birch leaves. Actually, this made me feel the need for more precise methods.

The solution was analysis by synthesis, and I first applied it to the singing voice, the most common of all music instruments. The method implies that you analyse the object by synthesising it. If you want to describe what characterises a singer's voice, you simply synthesise it. As soon as your synthesis contains all the timbral characteristics of the original, you know that from a perceptual point of view your synthesis is exhaustive. If the synthesiser is constructed as an analogue to the vocal apparatus, i.e., if it contains a set of formants attached to a voice source, just as the voice organ, your description is likely to be quite informative.

There are several important advantages with the analysis-by-synthesis method. 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 synthesis. A third advantage is that working with sound synthesis tends to draw your attention to details that may be quite important even though mostly unnoticed.

Correspondence concerning this article should be addressed to Johan Sundberg, Speech Music Hearing, KTH, SE-10044 Stockholm

Listening to the synthesis helps to direct your attention to such characteristics, and then, it is possible to define a terms for them.

Analysis-by-synthesis has been the main method in the development of the KTH system for synthesising singing. The well established acoustic theory of voice production was an important advantage, providing a solid starting point.

BACKGROUND AND ANECDOTES

MUSSE

Speech research has convincingly demonstrated that synthesis is a powerful tool in scientific research and Gunnar Fant and his Speech Transmission Laboratory, KTH had reached a leading international position in the area of speech analysis and synthesis. Therefore, singing synthesis was a natural thing to try at this department.

The start was the realization, in terms of a thesis work, of an idea of my department colleague Jan Gauffin to construct a hardware singing synthesizer, intended as a musical cousin to Gunnar Fant's classical speech synthesizer OVE ("Orator Vox Electrica"). One of Gauffin's main ideas was that formant frequencies and other synthesis variables should be continuously variable rather than variable in small but discrete steps. The idea was realized 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). Figure 1 shows a block scheme of the machine. It was played from a keyboard and provided with a number of interesting facilities related to the characteristics of classical singing. For example, apart from five formant frequencies and bandwidths, vibrato rate and extent, glottal noise, random variation of 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.

Possibilities to control MUSSE also by digital signals were created in terms of an interface, MUSSE DIG (Malmgren, 1978). A modified text-to-speech conversion system, RULSYS, developed within the department (Carlsson & Granström, 1975) was used to control MUSSE via the interface. This allowed the conversion of input music files into performances of *vocalises*, which are 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 working with the singing synthesis were revealing. It became evident that the MUSSE synthesizer could produce synthesis of excellent quality from the point of view of voice quality, but that the synthesis was poor 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. Attempts were made in collaboration with Rolf Carlson and Björn Granström in the speech group of the department to cure this deficiency by taking advantage of the context dependent rule tool that they had incorporated into their RULSYS program. By implementing context dependent accent, phrasing and marcato rules, the life-less character of a performance of a Vocalise by Panofka could be efficiently reduced, particularly when a live piano accompaniment was added. Indeed, when this synthesis was presented at an IRCAM symposium in Paris 1977 it triggered the only spontaneous applause of the audience at that conference. The dead-pan and the final versions of the Panofka Vocalise can be compared in **Sound Example 1**.

Encouraged by this success, I sent (with a faked name) the same recording of the final version of the Panofka Vocalise to one of the leading choirs in Stockholm, pretending that this was an example of my own singing and asking if the choir conductor would accept me as a member of his choir. After two weeks of hesitation, which I found extremely exciting and encouraging, the leader responded that all members of his choir were required to sing also consonants. This was a striking demonstration of MUSSE's limitation to vowel synthesis only. Possibilities to synthesize consonants were established some years later, also in terms of a thesis work (Ponteus, 1979). The first synthesis concerned the solmization syllables and was carried out mainly by Jan Zera, a Polish guest researcher (Zera et al., 1984).

Dynamic changes were modeled in MUSSE in terms of variation of the amplitude of the voice source signal. This produced a peculiar effect. Crescendos and diminuendos sounded as variation of microphone distance rather than as a variation of vocal loudness. The reason for this was that in reality an increase of vocal loudness is accompanied by a decrease of the source spectrum envelope slope. This effect was modelled by

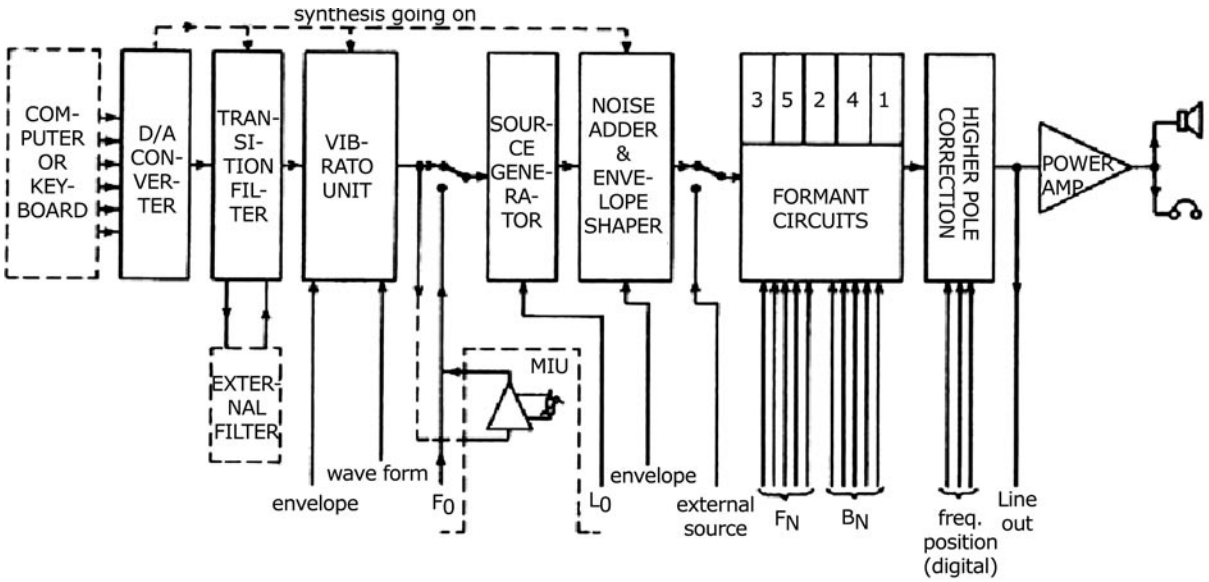

Figure 1.

Block scheme of the MUSSE singing synthesizer. The order of the formant circuits was designed so as to minimise noise in the system.

including in the system a "physiologic volume control", which was controlled by the same signals as that used for controlling the voice source amplitude.

The hardware MUSSE was replaced by a software version during the 90s. This implementation was realized by Sten Ternström.

The music performance research

In 1979 I started a cooperation with the violinist and music performance teacher professor Lars Frydén. This cooperation lasted for more than 20 years, until Frydén's death in the year 2000. The starting point for his interest may have been the question to what extent a musician could be replaced by a computer. A set of experiences formed the background of this question. One was his study under the Hungarian musician Istvan Ipolyi who formulated a set of simple rules for how music should be performed (Ipolyi, 1952). Another perhaps was his extensive teaching experience, where some basic instructions would tend to recur. In any event, Frydén was eager to find out to what extent Ipolyi's rules, complemented with rules that Frydén himself had tentatively formulated over the years, would appreciably improve the performance of music excerpts. The strategy was analysis by synthesis, i.e., to implement a performance rule into the RULSYS program and to apply it to various music excerpts.

The performance research, which took place during after-work-hours-sessions when Frydén was free from work and when the computer was free, yielded experiences that were extremely revealing. One striking observation was that we never disagreed whether or not a performance became better or worse by applying a specific rule, nor if the effect induced by a rule was exaggerated or appropriate. This negated the widespread assumption that the variability among musically acceptable performances of a given piece is unlimited. By contrast, our experiences demonstrated that music performance is restricted by a number of regularities that actually exclude most performances as being musically pathologic.

Another observation was that a great effect on a performance often could be achieved by applying a small number of rules or even one single rule. This seemed to suggest that the main message, which a listener apparently needs to receive from a performer, is that the performer cares about the piece of music being performed. It is tempting to speculate that this need is not specific to music communication but applies to communication in general; it typically is quite boring to listen also to an oral presenter who signals indifference regarding the message of the presentation.

Gradually a performance grammar emerged from these attempts to synthesize music performance. It was entirely derived from Lars Frydén's attempts to teach the computer program how to perform music in a musically acceptable way. This grammar, which can be seen as a scientific description of some aspects of Frydén's musical competence, is described and discussed by Anders Friberg in this volume. Here I will mention some examples of how synthesis of singing has efficiently helped develop a scientific understanding of vocal art.

The MUSSE singing synthesis was by no means unique. Rather, I repeatedly met people at conferences who told me they were working on singing synthesis. One prominent example was the Chant program, developed by Xavier Rodet at IRCAM in Paris. When I planned the second Stockholm Music Acoustic Conference SMAC II in 1993, I realized that it would be interesting to combine all these syntheses into one piece of music. I asked my friend Gerald Bennett, professor of composition at Hochschule Musik und Theater in Zürich and asked him if he would be willing to compose a piece of music for an International Ensemble of Synthesized Singers (IESS). He enthusiastically agreed and wrote the piece *Limericks*, with six verses, one for each of the centers invited to contribute. Unfortunately, Rodet could not participate as planned in this project because of time problems, so MUSSE had to perform also the verse intended for the Chant program. The music was processed by the performance grammar and the time table for the various notes was then distributed to the respective centers. The end result, provided with a synthesized piano accompaniment and mixed and edited by Anders Friberg and Sten Ternström had its world premiere at the closing session of SMAC 93. The synthesis strategies used for the various verses were quite different as can be noted by listening to the CD included in the Proceedings of SMAC 93 (Friberg et al, 1994).

Synthesizers represent a powerful tool not only in music performance but also in scientific research. The background is a consequence of the boundary between the physical world and the perceived world. With today's technology, it is reasonably easy to describe the physical properties of sound. For example, we can readily specify how the fundamental frequency, the amplitudes or the spectrum partials vary over time in a certain instrument. Yet, such descriptions may be called into question if they pretend to specify what is *typical* for that instrument. Such a claim can find support in synthesis work.

The KTH singing synthesis has been used in several applications and has been described in a handful of publications (Larsson 1977; Berndtsson & Sundberg, 1994; Berndtsson, 1995; Carlson et al, 1991; Sundberg, 1978b; 1989; 1981). In the present article, a few of these applications will be highlighted, the aim being to demonstrate how high-quality singing synthesis has been used in research in the area of music science.

SYNTHESIS RULES

Using the RULSYS method, rules were developed mainly during the 70s and early 80s for a number of musical contexts. Some of the resulting effects will be described here.

The rule *duration of consonants* takes into account the fact that in many languages consonant duration depends on length of the vowel preceding it. Thus, a long vowel is followed by short consonants and vice versa. This implies that the duration of a syllable will be different depending on if it is counted from the onset of the consonant or from the onset of the vowel, as illustrated in Figure 2. For example, say that the note should be 400 ms long. According to the rule system, a following consonant /// will be 0.5*400=200 ms if the vowel is short and 150 ms if the vowel is long, regardless of the note's duration. The principle applied in the MUSSE program is that consonants are considered part of the old note and that, consequently, all notes start with a vowel onset. The effect of this rule was illustrated in a sound example included in a previous article about the KTH singing synthesis system, but regrettably there was a timing error in the example. For this reason a corrected version of the same example is given in **Sound Example 2**. In both versions all notes have their nominal durations. In the first version, note duration is counted according to the orthographic principle, so that each tone starts with the consonant. In the second version each note is counted from vowel onset to vowel onset. Note that the seventh note in the example appears to arrive too early. This effect can be explained if it is assumed that note onset in singing is located at the vowel onset rather than at the consonant onset. If this principle is applied, the seventh note arrives too early because the sixth note is 88 ms too short.

By and large, the principle that tones start at the vowel onset in singing is in accordance with the result of an experiment where non-singer subjects were asked to pronounce syllables in a metrically regular sequence in synchrony with visual and acoustic timing cue appearing at a constant interval (Rapp, 1971). Generally, the subjects synchronized vowel onset with the timing cue, although vowel onset tended to lag behind the time marker in the consonant clusters /str/ and /st/. This delay may very well be due to the subjects' lack of training.

Timing of pitch change demonstrates a small but important detail of sung performance. The rule states that the pitch change should take place during the consonant preceding the vowel to be sung on the next note. Thus, the new note should begin with its target F0 rather than with an F0 that approaches this target. **Sound Example 3** presents an example of both cases.

Synthesis is a valuable tool for examining the perceptual relevance of various acoustic patterns and regularities, as mentioned. Thus, the *timbral consequences of a higher larynx* were analyzed by means of a listening test with synthesized ascending scales (Sundberg & Askenfelt, 1983). The aim was to assess the perceptual relevance of three acoustic consequen-ces that could be expected to accompany a rise of the larynx, (1) a small increase of the formant frequencies, (2) a decrease of the vibrato extent and (3) a decrease of the level of the voice source fundamental. The first change is a physical consequence of the shortening of the vocal tract which necessarily is associated with a rise of the larynx. The two latter changes were based on the hypothesis that a larynx rise is associated with firmer glottal adduction, resulting in a more pressed phonation type. Various versions of an ascending scale were presented to a panel of voice experts who were asked to rate how realistically the synthesis imitated a singer who raised his larynx with rising pitch. The results showed that the increase of formant frequencies was the most revealing characteristic of a rising larynx. This is not surprising, since the formant frequencies must increase when the larynx is raised. The decrease in vibrato extent and the decrease of the level of the fundamental were less efficient in producing the image of a singer raising his larynx with increasing pitch. **Sound Example 4** presents four versions of an ascending scale synthesized as specified in Table 1.

Table 1.

Organization of Sound example 4.

1. all parameters constant
2. amplitude of voice source fundamental decreases with pitch
3. same as 2, but also vibrato extent decreases with pitch
4. same as 3, but also formant frequencies increase with pitch

Diphthongs were realized by letting the formant frequencies of the first vowel remain for 35% of the tone's duration on the values belonging to the first vowel and then starting to approach the formant frequencies of the second vowel in the diphthong.

All singing voices except basses and perhaps baritones sometimes sing at fundamental frequencies which are higher than the normal value of the first formant frequency in at least some vowels. The situation that the frequency of the fundamental exceeds

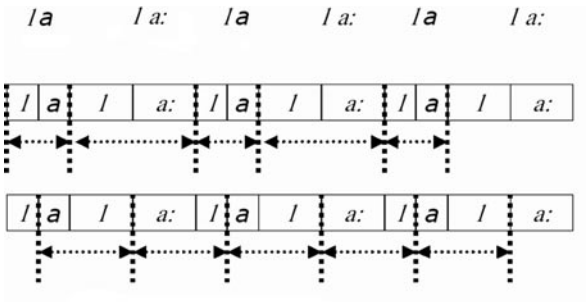

Figure 2.

Consequences for timing of definition of syllables. The figure shows the same sequences of the syllable /la/ shown at the top. In the middle row the syllable is defined as a consonant+vowel unit, in the bottom row it is defined as a vowel+consonant unit.

that of the first formant is typically avoided by classically trained singers by means of *formant tuning*, implying that the first formant frequency is increased such that it is not lower than the fundamental. This is the reason of the pitch dependent jaw and lip opening that typically can be observed when female vocalists sing at high pitches. The strategy which was observed already in the 1970s (Sundberg, 1975) was modeled in the synthesis program. It turned out that the voice quality sounded unnatural and shrill if the first formant frequency equaled that of the fundamental frequency. A better quality was obtained when the first formant was about 4 semitones higher than the fundamental. Later, the advantage of this strategy has been explained also from an acoustical point of view (Titze, 1994). Measurements on syntheses showed that this strategy increased the sound level of a vowel substantially, under some conditions by more than 10 dB.

Overtone singing is a special type of vocal art practiced in some Asian countries. It is characterized by the simultaneous appearance of two pitches in the sound produced by one single singer. It seemed reasonable to assume that this was a case of tuning formant frequencies such that they agreed with the frequency of a spectrum partial, which therefore was enhanced. Experiences from the MUSSE synthesizer showed that the effect was rather weak if only the second formant frequency was tuned to the frequency of the partial to be enhanced. Instead the second and third formants were tuned to a cluster with the second one placed on the frequency of the partial. Using this rule, the overtones of a constant drone tone could be made salient. In **Sound Example 5** the drone is shifted as soon as overtones of other tones are needed to form a melody.

Coloratura passages, i.e., rapid sequences of short notes sung on a single vowel need to be performed in

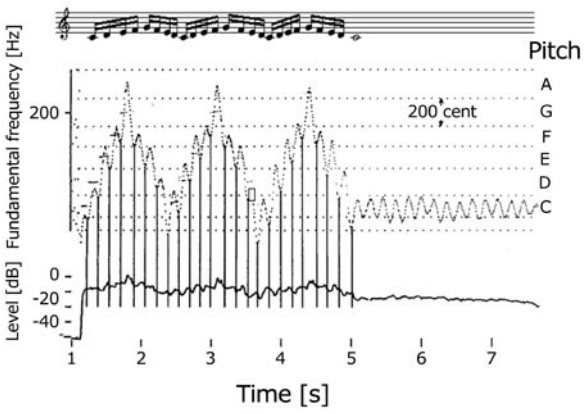

Figure 3.

Fundamental frequency pattern observed when a professional singer performed the coloratura sequence shown

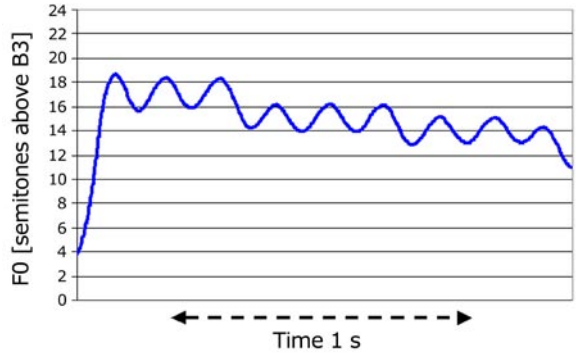

Figure 4.

Example of "Bull's Roaring Onset", i.e., a tone onset where the fundamental frequently starts about 12 semitones below te target which is then quickly reached, after about 100 ms.

a special way. Analysis of coloratura passages showed that singers make a turn with their F0 around the target F0 values, as illustrated in Figure 3. Thus, each target frequency is represented by a complete vibrato cycle. This apparently is produced by a pulsating subglottal pressure (Leanderson et al, 1987). An attempt was made to model the fundamental frequency pattern of coloratura passages (Sundberg 1981). The rule stated that in such contexts F0 should start 1 semitone above the target F0 at the onset of a tone, and that it should be one semitone below the same target in the middle of the note, and then, at the end of the tone, it should be one semitone higher than the next target F0. At the onset of the peak note of a melodic sequence that finishes an ascending series and initiates a descending sequence of notes, the initial F0 value had to be set to the target F0 + two semitones in order to produce the correct pitch. The F0 smoothing filter, which converted control voltage steps into cosine curves, produced a realistic synthesis of this type of singing. **Sound Example 6** presents two versions of a sequence of short notes, the first without and the second with the coloratura rule applied.

Bull's Roaring Onset was an attempt to model an F0 gesture frequently observed in sung performance. Figure 4 shows an example. At a phrase onset, F0 often can be seen to start at a low value and then to quickly to approach its target value. Similarly, F0 often drops considerably at the end of a phrase followed by a silent segment. In the MUSSE program this characteristic was produced by letting F0 of long notes start 11 semitones below the target F0, provided this F0 was higher than the pitch of C4.

THREE APPLICATIONS

Center frequency of the singer's formant

The singer's formant is a high spectrum envelope peak characterizing Western male operatic singing (Sundberg, 1987). A typical example of the use of synthesized singing was an investigation of the perceptual relevance of the center frequency of the singer's formant (Berndtsson & Sundberg, 1994).

The starting point was a striking experience made in synthesizing singing voices. The singer's formant can be generated by tuning formants number 3, 4, and 5 such that they come about 200 or 300 Hz apart from each other thus forming a cluster. This cluster is about 500 Hz wide, and partials falling into it become strong. As male singers produce a singer's formant in all voiced sounds, consonants as well as vowels, it produces a marked peak in a long-term-average spectrum (LTAS) of male operatic singing.

A typical experience in synthesizing singing voices was that a bass-like voice timbre was obtained when the singer's formant was centered around a lower frequency and a tenor-like timbre emerged when it centered around a somewhat higher frequency. An LTAS analysis confirmed this observation (Sundberg, 2001).

These results suggested that a bass, a baritone or a tenor voice timbre would emerge if the center frequency of the singer's formant was low, medium and high, respectively. This assumption was tested in an investigation where synthetic sung vowels were perceptually evaluated by a panel of experts. The stimuli consisted of descending triads that started from A3, D4, G4, or C5 (**Sound Example 7**). The panel's task was to classify the voice as bass, baritone, tenor, or alto. The classifications were assigned quantitative values, 0 for bass, 1 for baritone, 2 for tenor and 3 for alto. Using these values an average was computed for each stimulus. Figure 5 shows these mean ratings as function

of the start pitch of the stimulus triads. The figure shows that both the center frequency of the singer's formant and the pitch is relevant to voice classification. The influence of pitch seems a trivial result, since each classification has a typical pitch range. The relevance of the center frequency of the singer's formant is more interesting, since it corroborated informal observations made when synthesizing singing voices. It can also be observed that it would have been quite difficult to arrive at this result without the aid of synthesis. Singer's are not likely to obey if told to vary the center frequency of the singer's formant while keeping everything else constant.

Secrets of an ugly voice

Timbral beauty is an important aspect of a singer's voice, but certainly also elusive: what is beauty and from what does it emerge? Yet, the beauty of a voice is often striking, but perhaps, and even more so, the ugliness of some voices. A classical example of vocal ugliness has been published by the gramophone company RCA, "The Glory???? of the Human Voice". It contains the now classical recording of soprano Florence Foster Jenkins attempting to sing, among other things, the aria of the Queen of the Night from WA Mozart's *Zauberflöte*. It also presents a tenor-baritone who performs excerpts from Charles Gounod's opera *Faust*. The ugliness of his voice is quite monumental, as can be noted in **Sound Example 8**, which also presents, for comparison, the same excerpt as performed by the renowned Swedish tenor Nicolai Gedda.

The ugliness of this voice was eloquently commented upon in a review published on the Internet (http://www.epinions.com/content_84551175812).

"The disc concludes with something extraordinary, indeed. If you thought that Jenkins was bad, wait until you hear the selections from Gounod's Faust as sung by Jenny Williams (soprano) and Thomas Burns (baritone). Having translated the French text into English (a dubious endeavor), they proceed to out-do Jenkins in their awfulness. Actually, Williams is merely mediocre (i.e. a few notches above Jenkins). But Thomas Burns is extravagantly bad. In all truth, he sounds uncannily like Elmer Fudd, with the same nasal voice and portentous, tragic vibrato. Hearing his litany of "O! Marguerita"s and "I love you!"s belted in earnest, throaty groans is to witness the airy heights of absurdity...."

The ugliness thus described is quite fascinating, as it does not appear to be solely related to singing wildly out of tune, as in the case of the recordings of Florence Foster Jenkins. Rather it seems affiliated with the voice timbre. This raises the fascinating question why a voice timbre can be ugly *in itself*.

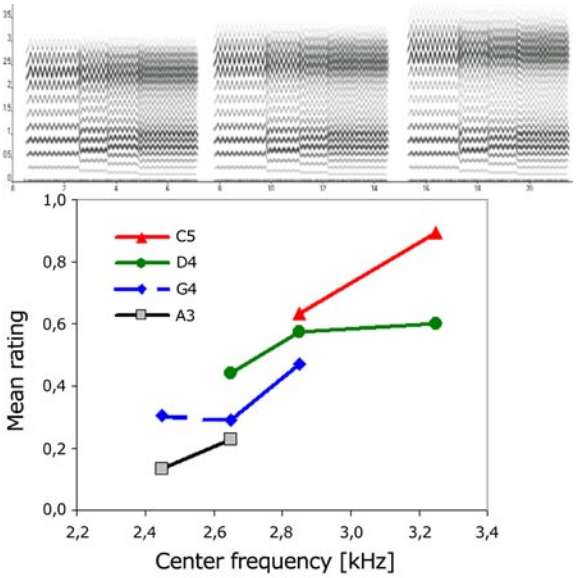

Figure 5.

Mean ratings of voice classification of descending triad stimuli starting from the pitches listed and with varied center frequency of the singer's formant. The classifications were quantified by assigning a value of 0, 0.33, 0.67 and 1.0 to votes for bass, baritone, tenor, or alto.

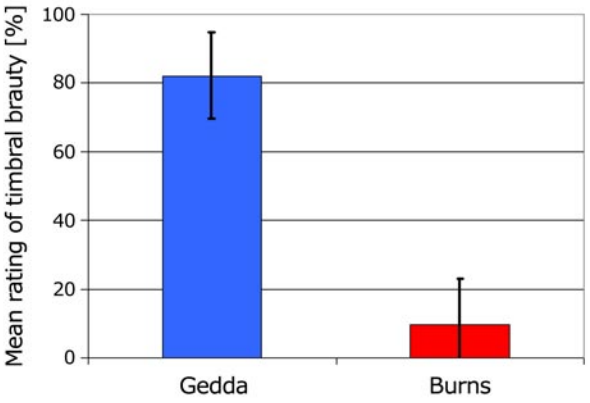

Figure 6.

Mean rating of timbre along a visual analog scale ranging from 0 (extremely ugly) to 100, (extremely beautiful) of Gedda's and Burns' voices by a panel of voice experts. Bars represent +- 1 SD.

Before embarking on an investigation of this, it seemed important to contemplate the subjectivity of ugliness, the main issue being to what extent experts of singing would agree about the ugliness of this voice. A panel of 15 singers or singing teachers were asked to rate, along a visual analogue scale, the timbral beauty of this excerpt as sung by Gedda and by the ugly voice. Each stimulus occurred twice in the test. The result came out as expected, as shown in Figure 6. Mean rating of the timbral beauty of Gedda's voice was 82 (SD 12.5) and that of Burns was 11 (SD 13.4), both out of 100.

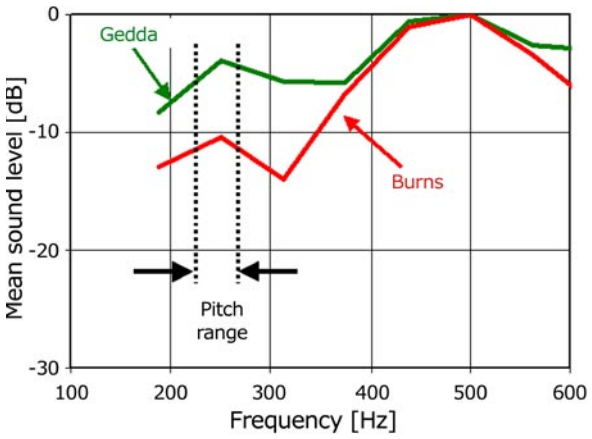

Figure 7.

Blow-up of low-frequency part of the LTAS of the solo parts of the excerpt analyzed as performed by Gedda and Burns (green and red curves, respectively, cf Sound Example 8). The dotted lines show the pitch range covered in the excerpt.

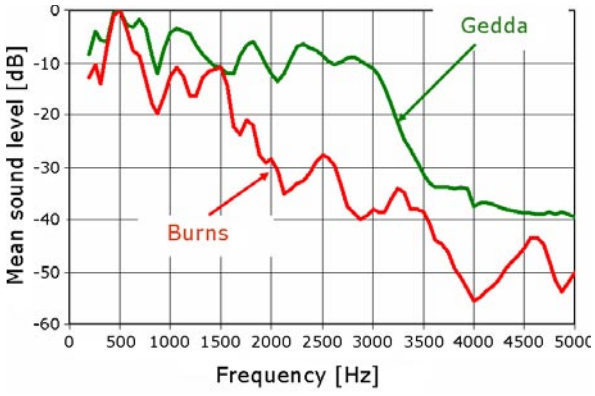

Figure 8.

LTAS of the solo parts of the excerpt analyzed as performed by Gedda and Burns (green and red curves, respectively, cf Sound Example 8).

It does not seem overly far-fetched to assume that ease of tone production characteristics contribute to the beauty of the voice. Studies of the glottal voice source have revealed a strong relationship between vocal pressedness or hyperfunction and the amplitude of the voice source fundamental (Gauffin & Sundberg, 1989). The background is that this amplitude is correlated with the peak-to-peak amplitude of the glottal airflow waveform, which, in turn, is closely related to the vibration amplitude of the vocal folds (Sundberg, 1995). An increase of glottal adduction tends to reduce this amplitude and hence attenuate the fundamental.

An LTAS offers a possibility to estimate the mean amplitude of the fundamental. Figure 7 shows the LTAS of the excerpt of both Gedda's and Burns' voices. The amplitude in the region of the fundamental in the excerpt analyzed is markedly lower in Burns' than in Gedda's LTAS. This difference may reflect a deficiency of the frequency curve of Burns' recording. On the other hand, such deficiencies were more typical of the high-frequency range than of the low-frequency range. This suggests that Burns was using a more hyperfunctional type of phonation than Gedda.

The singer's formant is a spectrum envelope peak in the range 2.5 to 3 kHz that belongs to the characteristics of male opera singers' voices. Typically it is present in all voiced sounds, consonants as well as vowels. For this reason, it is readily visible in an LTAS. Its center frequency varies between voice classifications, tenors tending to have a higher center frequency than basses. It helps the singer's voice to get heard against the background of a loud orchestral accompaniment, but it is also used when the singers sing with a piano accompaniment. Figure 8 shows the LTAS of Gedda and Burns. Again there is a striking difference. Gedda's voice has a clear singer's formant with a center in the vicinity of 2.7 kHz, Burns' voice, on the other hand, does not show anything similar to that. Rather his voice shows a marked peak near 3.25 kHz. A peak in this frequency range is characteristic of pop singers' voices (Cleveland et al, 2001). Thus, Burns lacks a singer's formant.

Vibrato is an important tone characteristic in operatic singing. Physically the vibrato is a slow, quasi-sinusoidal modulation of the fundamental frequency, the modulation frequency and amplitude mostly lying in the range of 5 to 7 Hz and 50 to 100 cents peak-to-peak. The perceived pitch of a vibrato tone corresponds to the frequency average (Sundberg 1978a, Shonle & Horan, 1980). This implies that the vibrato needs to be periodic in order to produce a constant pitch. Figure 9 shows the two recordings of the pitch F4. Gedda's vibrato is regular while Burns' is clearly irregular. This implies that the pitch perceived of Burns' voice is unstable. There is also a difference in the mean F0 during the tone, i e in the pitch perceived. The equally tempered value of this note is 8 semitones above the 220 Hz reference. Gedda's curve averages at 8.4 semitones, which is 0.4 semitones higher than the equally tempered reference, while Burns' average is 8.0 semitones. Thus, as compared with Gedda, Burns is flat on that tone. In addition, while Gedda's F0 remains constant throughout the tone, Burns' F0 descends by about 0.5 semitone toward the end of the note. It can also be observed that Burns gives a clear example of a "bull's roaring onset", starting the note with a wide and quick, ascending F0 glide.

These comparisons between Gedda's and Burns' voices revealed three clear differences, the mean amplitude of the fundamental, the intonation and the

singer's formant. Their perceptual significance was tested by synthesizing Burns' voice, including all its deficiencies, and then eliminating them, one by one (**Sound examples 9-13**). These synthesized examples were included in the same listening test as the one in which experts rated the timbral beauty of Burns' and Gedda's voices. The MUSSE DIG system was used for the synthesis.

The first version (**Sound examples 9**) reflects an attempt to replicate the main characteristics of Burns' voice. In the second version (**Sound examples 10**) the amplitude of the voice source fundamental was increased by a few dB. This would correspond to a reduction of the glottal adduction, i.e., slightly less hyperfunctional/pressed voice. The third version (**Sound example 11**) was to make the vibrato more periodic and to tune the intonation in accordance with the equally tempered tuning. The fourth version (**Sound example 12**) was to introduce a singer's formant by clustering formants 3, 4, and 5 in such a way that its center frequency appeared at about 2550 Hz. This is a value typical a baritone timbre, while the example was intended for a tenor voice. In the fifth version (**Sound example 13**), therefore the center frequency of the singer's formant was shifted 200 Hz up, to about 2.75 Hz.

The mean and SDs of the panel's ratings of the timbral beauty of these syntheses are shown in Figure 10. The mean rating of the first version that included all deficiencies noted in Burns' voice yielded a mean rating of 8.6 (SD 6.9) which is similar to that obtained for Burns' voice (9.6, SD 13.4). It can also be noted that each modification caused an increase of the rating mean value. Amplifying the fundamental and introducing a singer's formant increased the mean rating considerably, by almost 80% and 60%, respectively. The test results thus support the assumption that each of the observed deficiencies was relevant to the timbral beauty of the voice.

These results are interesting, since all of the reasons for timbral ugliness seem related to functionality. Lack of a singer's formant implies that the voice will fail to be difficult to hear against the background of a loud orchestral accompaniment. An irregular vibrato implies that the perceived pitch of a tone constantly varies so the pitch contour is not accurately realized. A constant use of hyperfunctional phonation is likely to limit the singer's range of timbral variation which would be needed for the purpose of expression. Against this background it is tempting to speculate that part of the criteria of timbral beauty in a singer's voice do not emerge from a randomly developing cultural tradition, but rather are rooted in the acoustic conditions under which singers create their vocal art.

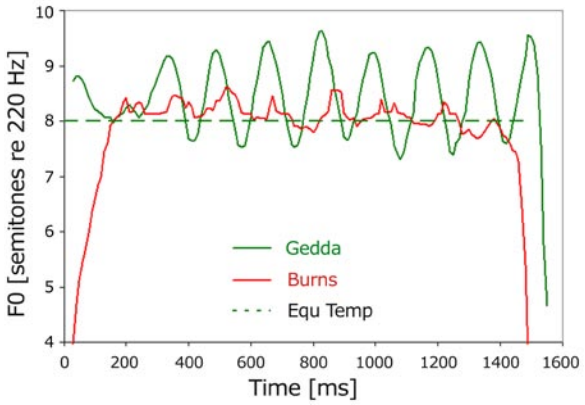

Figure 9.

Gedda's and Burns' intonation of the pitch of F#4 in the excerpt analyzed (green and red curves respectively). The dashed line represents equally tempered tuning.

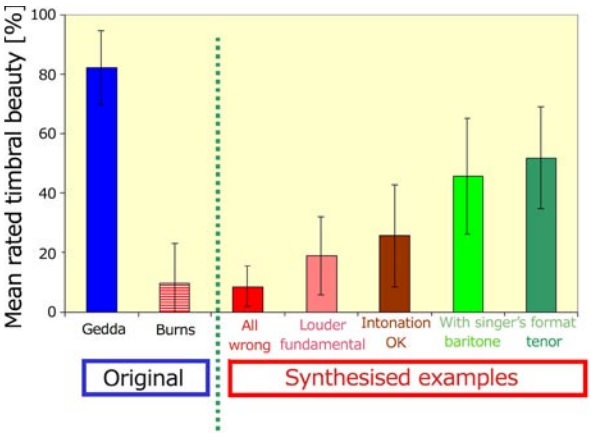

Figure 10.

The mean of the panel's ratings of the timbral beauty of the syntheses described in the text and given in Sound examples 9 – 13. Bars represent +- 1 SD.

Intonation

In keyboard instruments the fundamental frequency is predetermined for each key. On the other hand many other music instruments have free access to pitch variation, such that the player can decide the fundamental frequency within rather wide limits. The human voice is extreme in this respect. The limits of pitch variation are expanded also by the fact that vibrato is used in singing. Straight tones played on instruments that generate harmonic spectra produce beats in consonant intervals departing from just intonation. For example, a stretching of the pitch of C4 by 10 cent implies that its frequency is 1.5 Hz sharp, and this means that the tenth partial is 15 Hz sharp. Vibrato efficiently eliminates the risk of generating such beats, and that gives the singer access to intonation as an expressive means. This poses the question how singers take advantage of this freedom of intonation.

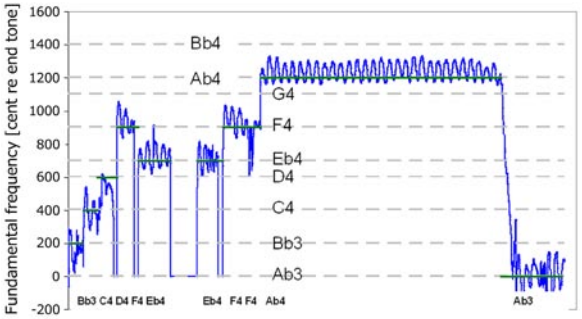

Figure 11.

Jussi Björling's intonation of a phrase taken from the Recitativo preceding Radames' Nile aria from G Verdi's Aida (cf Sound Example 14). The green lines represent equally tempered tuning.

In commenting on Figure 9 it was mentioned that Gedda's intonation of the pitch F4 is slightly sharp on this note, on average 42 cent. It is highly unlikely that this stretch is an example of singing out of tune. The excerpt was taken from a CD recording, which must have been scrutinized with extreme demands on intonation before it was published. Rather, the stretching must be assumed to belong to the expressiveness of Gedda's vocal art.

Further support for this assumption is provided in Figure 11 showing some other, similar examples of stretched intonation taken from the late Swedish tenor Jussi Björling's recording of G Verdi's *Aida*. Here, again, a stretching of the top note of a phrase by almost 70 cent can be observed. Thus in these two examples the intonation approached or even passed the category boundary between scale tones.

These observations pose the question how essential such positive departures from the theoretically correct intonation may be. An idea of this can be gained from **Sound Example 14**, where the top tone of the phrase appears in three versions. Example 14a presents Björling's original version. The top tone in this example was synthesized on the freeware formant synthesizer MADDE, developed by Svante Granqvist (http://www.speech.kth.se/music/downloads/smptool/). It allows control of formant frequencies and bandwidths, F0, vibrato rate and extent and source spectrum. These parameters were adjusted to produce a close match of Björling's tone, and then reverberation was added. Example 14b presents the result. In Example 14c the fundamental frequency of the top tone was lowered such that the top tone was exactly one theoretically pure octave above the final note. If special attention is paid to the pitch the difference is readily noticeable, and some listeners perceive a difference in expressivity between the examples 14b and 14c. This may or may not be the case, but in any event, intonation seems to be a relevant aspect of sung performance that calls for further investigation. A central question would be what tones should be sharp and what tones need to be in tune?

SOME FULL-FLEDGED EXAMPLES

Synthesis is a powerful in a singer's also rather subtle, though important aspects of singing. **Sound Example 15** is a synthesized choral tenor singer performing the theme of the first Kyrie of J S Bach's B minor Mass. The accompaniment, played on a real double-bass and a keyboard synthesizer, was added afterwards to the synthesis. The synthesis sounds quite realistic except for the upper tone of the ascending octave leap from F#3 to F#4 (370 Hz). The synthesis used the standard formant frequencies for the vowel /y/ at 330 Hz, which thus is lower than the fundamental of the F#. A real singer would prefer always having the first formant higher than the fundamental. The effect of not applying this principle can be heard as a queer tone quality on this particular note.

Another example of syntheses of non-operatic singing can be listened to in **Sound Examples 16** the first phrase of the *Pie Jesu* movement of Gabriel Fauré's *Requiem*. The solo voice is intended for a boy soprano. This synthesis was made with the formant frequencies of an adult male choir voice. These formant frequencies were then multiplied with a factor of 1.5, and the pitch was adjusted according to the score. The synthesis parameters were generated by the MUSSE program and then Sten Ternström used the parameter file to control the Aladdin signal processor system (http://www.hitech.se/develop11ment/) adjusted to copy the MUSSE synthesizer.

Sound Examples 17, the first movement of Claudio Monteverdi's *Maria Vesper*. gives another example of non-classical singing. All voices in this examples were synthesized on the MUSSE system. For the choral part, however, different formant frequencies were used in accordance with the observations that voices with a higher pitch range use higher formant frequencies than voices with a lower pitch range. The orchestral parts were synthesized by Anders Friberg using the Director Musices system (http://www.speech.kth.se/music/performance/).

A synthesizer does not suffer from the limitations of a real singer. This was taken advantage of by Gerald Bennett in the last verse of his SMAC 93 composition *Limericks*, **Sound Examples 18**. Here the solo voice is written for a synthesized singer of undetermined classification and with a pitch range that widely

exceeds that of any living singer. From D2 (73.4 Hz) up to B6 (1976 Hz). After the top note the singer performs a 52 semitones descending leap to the pitch of G2. In this example is also included the complete IESS singing the refrain.

DISCUSSION AND CONCLUDING REMARKS

Singing as well as most other acoustic signals are quite complex, and the acoustic properties of singing are particularly important, since vocal artists use them for constructing acoustic pieces of art. The task of music science is to answer the questions How? and Why?, i.e., to describe and explain music. Acoustic analyses tend to supply an overwhelmingly great amount of data, and the problem is to identify which ones are perceptually relevant. In this endeavor synthesis is an irreplaceable tool.

Here some examples of such work have been described. The acoustic characteristics of a larynx rising with pitch showed that the most reliable sign is that the formant frequencies increase. This seems logical, since, by necessity, a shorter vocal tract will have higher formant frequencies than a longer vocal tract. On the other hand, it would be possible for a singer to hide at least some of these effects by compensatory articulatory means. In addition, the experiment showed that also lower amplitude of the voice source fundamental and smaller vibrato extent were typically associated with a larynx rising with pitch.

Synthesis is also an indispensable tool for exploring the perceptual significance of a particular acoustic property. The study of the ugly voice is an example. According to the results of the listening test, the elimination of acoustic properties, suspected to belong to the acoustic description of timbral ugliness, actually increased the timbral beauty of the synthesis. The listening test also showed that the observations made were insufficient as a description of timbral beauty, since the mean rating of the best version was still far away from the mean rating of Gedda's voice. Thus, in this case synthesis could also shed some light on the question how much of timbral beauty and ugliness did the acoustic data explain? In addition the results invited to some speculation on the reason why ugly is considered ugly and beautiful is regarded as beautiful. In classical singing some of the timbral beauty seems related to functionality.

A condition for the usefulness of synthesis for these purposes would be that it possesses a high degree of naturalness. A striking unnaturalness of the stimuli in a listening test is likely to catch the listeners' attention and distract from the acoustic property to be analyzed. In such cases the response of the test is likely to be polluted by a great portion of random variation. Listeners would need to envisage a real singer behind the stimuli in order for them to rely on their experience of singing voices when responding

Synthesis of singing is also a powerful tool for directing a listener's attention to a specific aspect of a performance. To many listeners, the timing of pitch change is an example of this. Many listeners need repeated listening to the contrast between the two versions of the triad in Sound Example 3 in order to be able to identify and define the problem. This would imply that synthesis of singing has a great potential for voice pedagogy.

With regard to vocal art the goal of music science is to describe and explain what characterizes musically interesting performances. Indeed, this goal would hardly be possible to reach without tools that produce realistic synthesis.

Sound examples

Sound example 1.

Dead-pan and final versions of a Vocalise from Heinrich Panofka's The Art Of Singing, op.81. (From Sundberg, 1978)

Sound example 2.

Syllable boundary. Two versions of a theme from W.A. Mozart's opera Don Giovanni, where the duration of the consonant /l/ is greater after a short/unstressed vowel and shorter after a long/stressed vowel. The syllable /la/ is repeated for each tone, and in the first version, the syllable is counted from consonant to consonant, as in orthography, while in the second version the syllable is counted from vowel onset to vowel onset. The perceived rhythmical structure is affected by this difference.

Sound example 3.

Timing of pitch change. In the first example, the pitch change happens during the first part of the vowel, in the second version it happens during the consonant. Each version is played three times. (From Sundberg, 1989)

Sound example 4.

Larynx rising with pitch. Four versions of an ascending scale. In the first all parameters are kept constant throughout the scale. In the second the amplitude

of the voice source fundamental decreases gradually with increasing pitch in the final tones of the scale. In the third, the same is true also for the amplitude of the vibrato. In the fourth the same is true also for the formant frequencies.

Sound example 5.

Formant tuning. In this example, the second and third formants are tuned closed together near the frequency of a spectrum partial of a drone tone. Hence this partial becomes perceptually salient. By shifting the drone all scale tones of the major diatonic scale can be produced. The enhanced partials thus play the tune "Jinglebells". (From Sundberg, 1989)

Sound example 6.

Coloratura. Two versions of a sequence of short notes, the first without and the second with the coloratura rule applied. This rule forces the fundamental frequency to make a turn around each target frequency, as illustrated in Figure 4. (From Sundberg, 1989)

Sound example 7.

Center frequency of singer's formant. Descending triads starting from the pitch of D4. In the first one formant frequencies F3, F4, and F5 are at 2.2, 2.4, and 2.7 kHz, in the second one at 2.4, 2.6, and 2.9 kHz, and in the third one at 2.6, 2.8, and 3.1 kHz, respectively.

Sound example 8.

Timbal ugliness. An excerpt from Charles Gounod's opera Faust performed by Nicolai Gedda and by Thomas Burns

Sound example 9.

Synthesis of Burns' voice.

Sound example 10.

Same as Example 9 except that the amplitude of the voice source fundamental was increased by 4 dB.

Sound example 11.

Same as Example 10 except that the random variation of vibrato rate and extent was eliminated and that the fundamental frequency was tuned according to the equally tempered tuning.

Sound example 12.

Same as Example 11 except that the formant frequencies number 3, 4, and 5 were clustered, such that a singer's formant was created with center frequency at 2.5 kHz, approximately.

Sound example 13.

Same as Example 12 except that the center frequency of the singer's formant was increased by 200 Hz.

Sound example 14.

Three versions of an excerpt from Giuseppe Verdi's opera Aida performed by Jussi Björling. In the first version Björling performs the entire excerpt, in the second tone the top note was synthesized and in the third version the mean fundamental frequency of the top note was lowered by about 65 cent such that it agrees with the equally tempered tuning.

Sound example 15.

Synthesis of the first theme of the first Kyrie from J S Bach's B Minor Mass. The accompaniment was played on a doubhle bass and a keyboard synthesizer and edited into the file.

Sound examp le 16.

Synthesis of the first phrase, composed for a boy soprano, of the Pie Jesu movement from G Fauré's Requiem. The accompaniment was synthesized and edited into the file.

Sound example 17.

Synthesis of the first movement of C Monteverdi's Maria Vesper. Voices were synthesized by means of the MUSSE machine and the instrumental parts on a computer controlled synthesizer. (From Sundberg, 1989)

Sound example 18.

Last verse with refrain from G Bennett's Limericks, composed for the International Ensemble of Synthesized Singers (IESS). The voice was synthesized by means of the MUSSE machine and the piano on a computer controlled synthesizer. (From Friberg & al, 1994).

References

Berndtsson, G. & Sundberg, J. (1994), The MUSSE DIG singing synthesis. In A Friberg, J Iwarsson, E Jansson & J Sundberg (eds.), *SMAC 93 (Proceedings of the Stockholm Music Acoustics Conference 1993)*, Stockholm, Roy. Sw. Academy of Music, Publ. No 79, (pp. 279-281).

Berndtsson, G. (1995). *Systems for synthesizing singing and for enhancing the acoustics of music rooms*. Dissertation, KTH, Department of Speech communica-

tion and Music Acoustics, Royal Institute of Technology, Stockholm.

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.

Carlson, G., Ternström, S., Sundberg, J. & Ungvary, T. (1991). A new digital system for singing synthesis allowing expressive control. *Proceedings of the International Computer Music Conference*, (pp. 315-318). Montreal,

Cleveland, T., Sundberg, J. & Stone, R.E. (2001). Longterm-average spectrum characteristics of country singers during speaking and singing. *J. Voice, 15*, 54-60.

Friberg, A., Iwarsson, J., Jansson, E., Sundberg, J., (eds.), (1994). *SMAC 93 (Proceedings of the Stockholm Music Acoustics Conference 1993)*, Stockholm, Roy. Sw. Academy of Music, Publ. No 79.

Gauffin, J. & Sundberg, J. (1989). Spectral correlates of glottal voice source waveform characteristics. *J. Speech and Hearing Res 32*, 556-565. WWW

Ipolyi, I (1952). *Innföring i musikspråkets opprinnelse og struktur*, Bergen, Norway: JW Eides Forlag.

Larsson, B. (1977). Music and singing synthesis equipment (MUSSE). *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, 1/1977, 38-40.

Leanderson ,R., Sundberg, J., von Euler, C. (1987). Role of the diaphragmatic activity during singing: a study of transdiaphragmatic pressures. *J. Appl. Physiol., 62*, 259-270. WWW

Malmgren, J. (1978). *MUSSE Interface Unit MIU*. Thesis work, Department of Speech Music Hearing, KTH.

Ponteus, J. (1979). *Mimmi, en utrustning för konsonantsyntes avsedd att komplettera MUSSE*. (Mimmi, an equipment for consonant synthesis intended as a complement for MUSSE), in Swedish, Thesis, Department of Speech Music Hearing, KTH.

Rapp, K. (1971). A study of syllable timing. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, 1/1971, 14-19

Shonle. J. I. & Horan, K. E. (1980). The pitch of vibrato tones. *J. Acoust. Soc. Amer., 67*, 246-252. WWW

Sundberg, J. (1975). Formant technique in a professional female singer, *Acustica, 32*, 89-96. WWW

Sundberg, J. (1978a). Effects of the vibrato and the 'singing formant' on pitch. In *Musica Slovaca VI*, (Bratislava) 51-69.

Sundberg, J. (1978b). Synthesis of singing. *Sw. J. Musicol. 60*, 107-112.

Sundberg, J. (1981). Synthesis of singing. In R. Favaro (ed.), *Musica e Tecnologia: Industria e Cultura per lo Sviluppo del Mezzagiorno, VI Colloquio di Informatica Musicale, Symposium Proceedings*, (pp. 145-162). Edizione Unicopli.

Sundberg, J. (1987). *The Science of the Singing Voice*, Northern Illinois University Press, Dekalb, IL.

Sundberg, J. (1989). Synthesis of singing by rule. In M. Mathews & J. Pierce (eds.), *Current Directions in Computer Music Research* (pp. 45-55 and 401-403), System Development Foundation Benchmark Series, Cambridge, MA: The MIT Press.

Sundberg, J. (1995). Vocal fold vibration patterns and modes of phonation. *Folia Phoniatrica, 9*, 20-26 WWW

Sundberg, J. (2001). Level and center frequency of the singer´s formant. *J. Voice, 15*, 176-186.

Sundberg, J. & Askenfelt, A. (1983). Larynx height and voice source: a relationship?. In J Abbs & D Bless (eds.), *Voice Physiology*, Houston TX: Collegehill, 307-316.

Titze, I. R. (1994). *Principles of Voice Production*, Prentice-Hall: Englewood Cliffs, NJ.

Zera, J., Gauffin, J., and Sundberg, J. (1984). Synthesis of Selected VCV-Syllables in Singing. *Proc. International Computer Music Conference*, IRCAM, Paris, 83-86.