Proc.1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002), Leuven, Belgium, November 15, 2002

SYNTHESIS AND PROCESSING OF THE SINGING VOICE

Xavier Rodet

IRCAM
1, place I. Stravinsky, 75004, Paris, France
rod@ircam.fr

ABSTRACT

As soon as the beginning of the 60 sthe singing voice have been synthesized by computer. Since these first experiments, the musical and natural quality of singing voice synthesis has largely improved and high quality commercial applications can be envisioned for a near future. This talk gives an overview of synthesis methods, control strategies and research in this field. Future challenges include synthesizer models improvements, automatic estimation of model parameter values from recordings, learning techniques for automatic rule construction and gaining a better understanding of the technical, acoustical and interpretive aspects of the singing voice

1. INTRODUCTION

What is meant here by the terms "Synthesis and Processing of the Singing Voice" is the production of a human-like singing voice by a computer. The input data for the synthesis program can be a score in some standard format, such as the Common Music Notation, and lyrics given as ordinary text or phonetic notation. But, in another case, the input data can be a given *performance* also represented in some digital format, the performance of a human singer or an instrumental performance. In this case of a given *performance*, the aim of the synthesis program is to produce a human-like singing voice which mimics the given performance. Singing voice coding will not be addressed in this paper, even though structured encoding [39] is clearly related to synthesis.

Singing Voice Synthesis (SVS) has been a subject of research for more than 40 years, and the quality which can be obtained now opens new perspectives. Recently, SVS programs have been made available on the Web or as commercial products. Even though their singing voice quality is not that of a human singer, the new research results and expected progress are such that real commercial usage can be envisioned for a near future.

The structure of this paper is as follows. The history and detailed technical review of SVS will not appear in this paper, since they can be found in [11]. Section 2 lists various possible applications of SVS. In section 3 are presented some problems related to voice quality and aesthetics which are even more crucial for musical applications than for speech synthesis. The architecture of SVS systems and the techniques used by these systems for sound signal calculation, i.e. waveform *synthesizers* or *unit modification and concatenation,* are explained in section 4. Some of the recent advances in these techniques are exposed in sections 5. In order to render a score into a high quality singing voice, various levels of rules are required and these rules are listed in section 6. For singing synthesis, as for speech synthesis, large improvements have been obtained by the technique called *unit selection*. This technique is presented in section 7 and problems in the elaboration of the database of units are exposed in the next section. Some examples of recent SVS are listed in section 9. Choir singing and the difficult question of the evaluation of the SVS results are briefly presented in section 10 and 11. Finally the most important points to improve and some future research and developments are dealt with in section 12.

2. APPLICATIONS

A first class of applications is the one where SVSs have the same role as a human singer interpreting a score. This is of interest for music creation when a singer with the proper musical competence is not available, e.g. for non-singer musicians and amateurs. It can also be useful for post-processing in a music studio when corrections to a singing recording are required and the singer can not perform another recording. For instance, restoration of singing archives could be done by using a SVS system. As another example, new scores or new interpretations could also be created as if performed by a singer of the past or at least with a great resemblance to such a singer, or to his style. The three previous examples necessitate that a model of a given singer has been built and we will see that this can be done, from recordings, by a learning procedure.

In the previous applications, the input to the SVS is score and lyrics. In another case, the input data can be a given *performance* also represented in some digital format. In this later case, the aim of the synthesis program is to produce a human-like singing voice which mimics the given performance. A typical application is the voice conversion from one singer to another such as for karaoke. A second class of applications uses an instrumental performance as input to control features among which note duration, pitch, vibrato, dynamics, etc.. This is done for

example by [34], [35] where the instrumental performance can be from a violin.

Singing synthesis is useful in music creation to produce voices that a human would not be able to do. For instance, composer G. Bennett in his piece *Winter* [2] has used singing synthesis. Another example is the creation of a castrato voice for the movie *Farinelli* of G. Corbiaud. Few or no singer today is able to sing the castrato repertoire with the required voice and artistic quality. Therefore, a voice was created by our Analysis-Synthesis team at Ircam [14] by recording a coloratura soprano and a counter tenor, and by morphing their voices into to a single new voice supposed to have the characteristics of a castrato, and having the artistic qualities wished by the movie team.

Finally, it should be noted, as said by J. Sundberg [42], that acoustic analysis yields an abundance of perceptually irrelevant information. Therefore, synthesis appears as a forceful tool to check acoustic hypothesis. Sundberg has applied singing voice synthesis to characterize some of the most important features which distinguish good and bad quality singing. These features will be discussed in the next section.

Before the main technical discussions in the following sections, it is worth indicating that, for simplicity, the term *pitch*, i.e. a notion of perception, is often misused here. To be exact, it should often be replaced by *fundamental frequency*. However it is true that the perceived pitch of a singing voice is most of the time directly connected to the fundamental frequency of the corresponding signal.

3. PROPERTIES, QUALITY, VOCAL AESTHETICS

The properties of the voice which are aimed at in SVS are even more important than in speech synthesis. They are also referred to by words such as *quality*, *naturalness* and *vocal aesthetics*, to which a *musical* quality must probably be added [79]. However, even though they surely describe important properties, these words are rarely defined with precision. It has been suggested that naturalness could be defined so that listeners think a human singer produced the voice. In a case where the voice of a precise singer is the goal, one could as well add that the system's voice is recognized as possibly from *this* singer, i.e. the singer whose voice was used for system training. As an example about naturalness, one can listen to the sound example 1. available in

`http://www.ircam.fr/anasyn/reine.html`

It is a synthetic interpretation of the aria of the *Queen of the Night* in Mozart's Opera *The Magic Flute* which was done by Y. Potard and X. Rodet, [57], [2]. It has been seen several times that unaware listeners attribute the synthetic voice to a human singer and not to a computer and that they spontaneously emit judgements about the quality of the singer. Finally, an interesting approach of naturalness has been proposed by S. Tenrstr m in terms of a *layered transport model* of a form of communication [79].

In [73] some features which distinguish a good quality from a bad one have been tested by synthesis and listening experiments. The starting voice, considered as ugly, has little singer's formant, a relatively weak fundamental (so that the voice sounds *pressed*), an irregular vibrato and an unstable average pitch. Correcting the voice for a strong singer's formant, a strong fundamental, a regular vibrato and a stable average pitch improve largely the estimated quality. But when these features are valid for western classical singing, Sundberg notes that *pressed, no singer's formant* and *no vibrato* may well be observed in *pop* and *rock* voices considered to be good voices. This is true also in singing voices of non-western cultures. The feature which seems to remain valid through different styles seems to be a form of regularity versus irregularity. Clearly a lot more research is needed in the question of voice quality.

As a human can, a SVS system should also be able to produce different interpretations of the same piece. Expressivity as well is an essential quality of a musical system. Only few research works have been undertaken to understand and objectively characterize expressivity [16], [37]. For instance, emotion in music has been considered by Bresin and Friberg in [8], [27] and [18], and perception of emotions in singing is investigated in [33].

4. ARCHITECTURE OF SINGING SYNTHESIS SYSTEMS

As explained in section 2, SVS systems take as input a score and lyrics and, eventually an interpretation (Fig. 1). The score defines the notes to be sung, their mean pitch, their durations and some information about loudness (known as *dynamics*) and about attacks, note shapes and transitions between notes (known as *articulation*). To avoid confusion, this is named *musical* articulation here, as opposed to the articulation of the vocal apparatus named here *physical* articulation. In the case where the performance is not given, a set of performance rules is in charge of calculating it. More precisely these rules determine the characteristics of the vibrato [51], [72], [52], [80] and the exact value of the pitch at any instant. They also compute the precise loudness shape over time of each note and how successive notes are connected or not, and this is specified in terms of energy and some attributes of timbre such as spectral slope. All these values (vibrato, pitch, energy, timbre, etc.) are functions of the time and are often called *parameters*. Another set of parameters is deduced from the lyrics. From the lyrics orthographic text, phonetic rules deduce the phonetic text to be sung.

At this point, it is necessary to distinguish several classes of systems. The first distinction is between systems which use a (waveform) synthesizer (Fig. 1) and those which use the modification and concatenation of recorded units of natural singing voice (Fig. 2). In the later case, named *concatenative* synthesis, the parameters issued from the rules are used to choose, according to the phonemes to be sung, the best units in a database. Then these units, i.e.

their sound signals, are modified in order to obtain the expected properties (vibrato, pitch, energy, timbre, etc.).

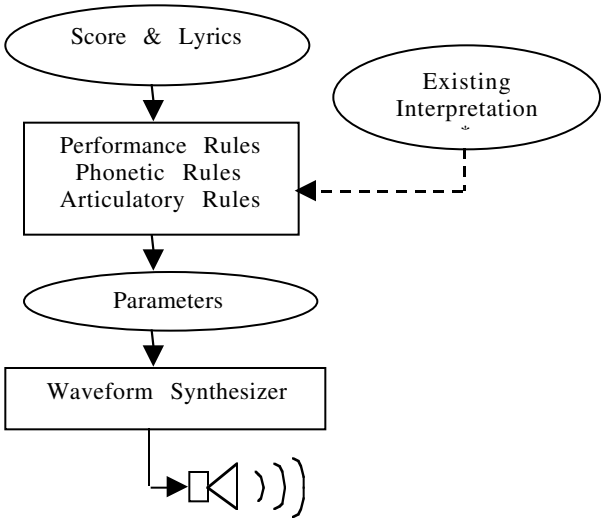

Fig. 1. System using a synthesizer.

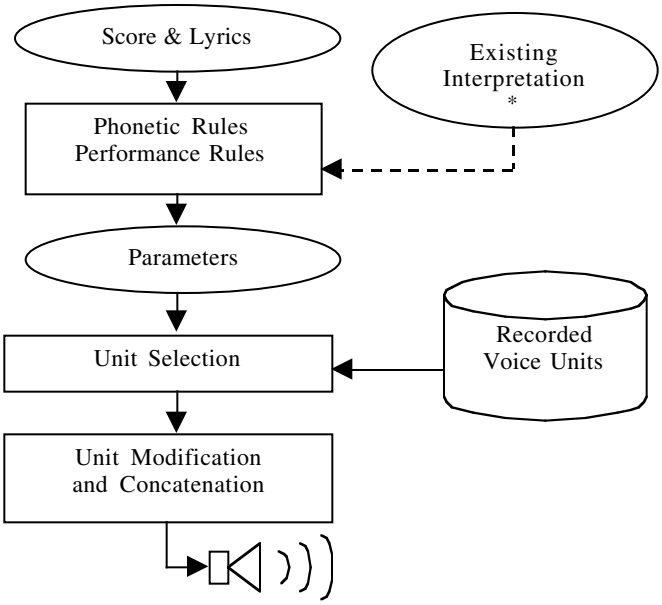

Fig. 2. System using modification of recordings.

In the case of a waveform synthesizer, another distinction is between synthesizers which are physical models of the voice production mechanism and synthesizers which are models of the voice signal. In the case of a physical model (also known as an articulatory synthesizer), *articulatory* (or *segmental*) rules transform the sequence of phonemes, and the other parameters, into parameters for the synthesizer, i.e. movements of the vocal apparatus elements, such as the tongue, the jaw, the vocal tract shape, the tension of the vocal folds, the sub-glottal air pressure, etc.. In the case of a signal model, the sequence of phonemes and the other parameters are used to compute the parameters of the signal model such as pitch, glottal pulse spectral shape, formant frequencies, formant bandwidths, etc.. Each of the different classes of systems described above present specific advantages and drawbacks.

The best systems therefore combine several approaches in order to improve the resulting singing voice. As an example, instead of a true physical model of the vocal folds, a glottal pulse shape can be used as the excitation of a model of the vocal tract [11]. In the same idea, a system can combine a waveform synthesizer and recorded voice units[6].

5. WAVEFORM SYNTHESIZERS AND UNIT MODIFICATION TECHNIQUES

The different sorts of waveform synthesizers will not be described here since they are given in [11]. But recent improvements in waveform synthesis quality have been obtained in source filter-models and also by a combination of the unit modification technique, signal models, as well as algorithms and rules deduced from physical models. Some of these advances are indicated in the present section. Source-filter models consider that voice production can be represented as a glottal excitation waveform [74] filtered by a linear filter. Aside the linear assumption, this model does not take explicitly into account the coupling between the glottis and the vocal tract.

For such a model, recent improvements provide a high quality analysis-synthesis method and a flexible control of vocal textures [32]. The model is composed of a transformed Liljencrants-Fant [22], [23] parametric source model and an all-pole filter. For each pitch period, the analysis procedure is an iterative simultaneous estimation of the glottal and vocal tract parameters. Several of the best solutions are retained since different source waveforms associated to different filters can produce nearly as good voice waveform estimates. Then a unique solution for each period is retained by use of a *Dynamic Time Warping* algorithm, which enforces continuity from period to period. In [31], the dependencies between the glottal waveform characteristics and the qualities of the voice have been studied. The glottal waveform is characterized by parameters known as the *open quotient*, the *asymmetry coefficient* and the *return phase quotient*. A model of the glottal waveform with these parameters has been implemented and tested in experiments of the perception of the synthetic singing voice. The results of this study allow the control of the qualities of a synthetic voice such as *pressed*, *normal* and *breathy*.

Sinusoidal plus Residual signal models [55] have been used with success to allow for the modification of pitch and duration of units. The SVS system by [42] introduces a technique, named SM-PSOLA, which combines Sinusoidal

Modeling and PSOLA (Pitch Synchronous Overlap-Add [10]). Improvements of the method named ABS/OLA (Analysis-by-Synthesis/Overlap-Add [29]) is proposed for singing voice synthesis in [40]). [62], [63], [59] suggest to concatenate transition and consonant units coded with the Sinusoidal plus Residual model with the high quality vowels obtained with the *Chant* model (Example 1. in section 3.) and coded in the same model.

One of the first sinusoidal models, SMS [65], is used in [6] where several other waveform synthesis novelties are implemented. Glottal excitation unit pulses are first built in the time-domain and transformed in the frequency domain where a filter is applied to form a glottal excitation in the frequency domain. Then a model of the vocal tract filter, in terms of its resonances, is applied to this glottal excitation in the frequency domain. Finally, amplitude and phase of sinusoids are computed for additive synthesis. This design gives a great flexibility allowing modifying at will many features of the synthetic signal.

6. RULES

Several different sets of rules are needed in order to convert score and lyrics into the parameters leading to an acceptable singing voice. The rules, which transform the lyrics into phonetics, have been already mentioned above, as well as the *articulatory* (or *segmental*) rules, which convert the sequence of phonemes into physical or acoustical articulatory parameters to correctly render the pronunciation of the text. The precise alignment of the phones with the notes has also to be determined [60][Ross & Sundberg 01]. A score only defines the notes to be sung, their mean pitch, their standard duration according to the tempo and some information about *dynamic* (loudness) and about *musical articulation* (attacks, note shapes and transitions between notes). But for a good singing voice performance, several acoustic characteristics and their time evolution have to be precisely calculated. One can distinguish *low-level* rules necessary for the perception of a human-like singing voice and *high-level* or *performance* rules necessary for a really musical performance. The main characteristics which have to be determined are the tempo, the deviation of note duration from the standard value, the average pitch over a vibrato period [17], the vibrato rate and excursion, the exact loudness and how successive notes are connected (or not) in terms of energy (e.g. legato, staccato, etc.) and some attributes of timbre such as spectral slope, formant frequencies and singer s formant. For each note and from note to note, a precise evolution in time of some of the characteristics has to be calculated. Application of low-level rules can be heard in example 1 in section 3 and observed in Fig. 3. From one note to the next the discontinuity in fundamental frequency is smoothed

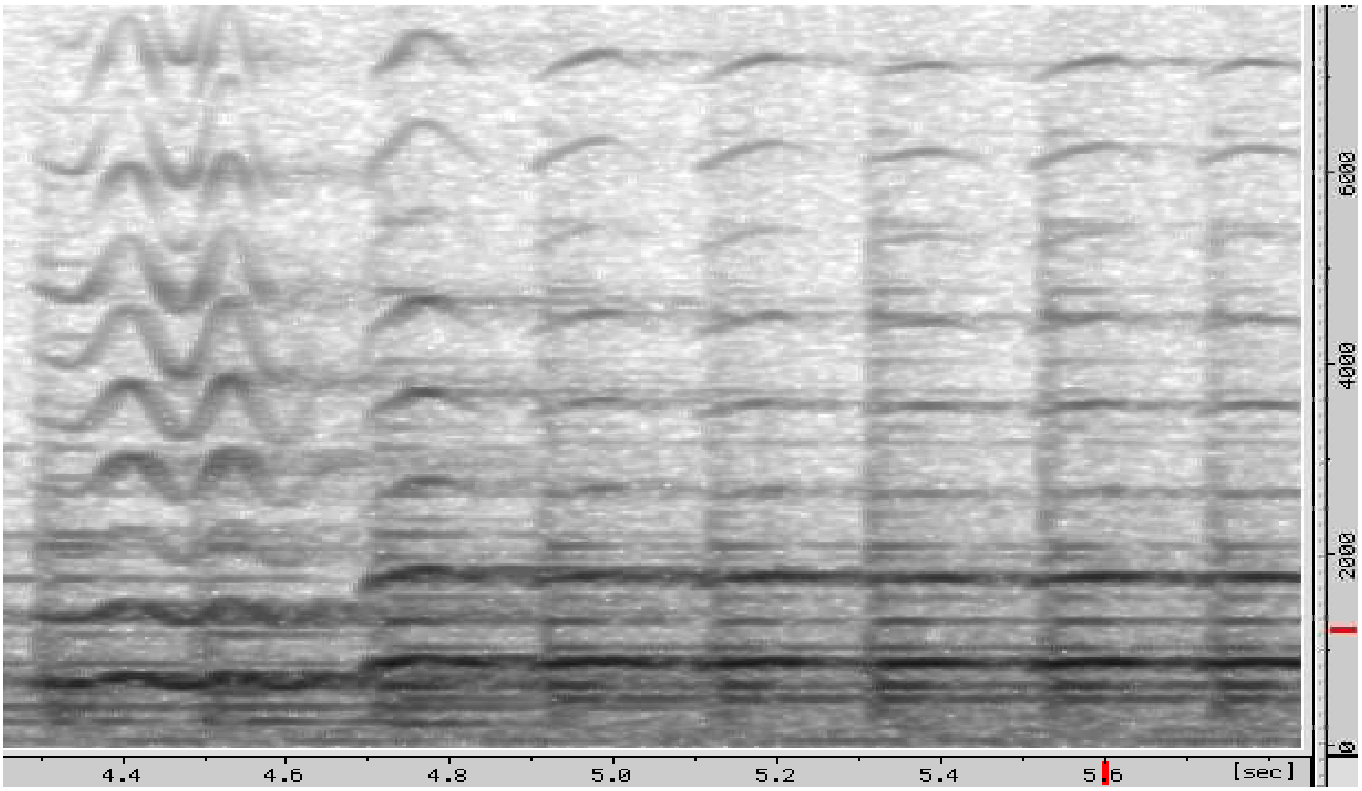

Fig. 3. Sonagram of an excerpt of synthetic example 1, section 3.

according to the interval. Also, in this example, for each note the average pitch usually starts below the standard pitch, has a maximum above it and ends below it. The vibrato rate, the vibrato excursion and the loudness follow a similar shape.

A rather large set of performance rules have been developed by J. Sundberg, A. Friberg and L. Fryd n[26], [71], [4] and [28]. According to the musical context of the notes, these rules vary tempo, duration, loudness, pitch and other features. Three types of rules are distinguished. The *differentiation* rules enhance the differences between successive musical elements such as tone categories. The *grouping* rules delimit the notes which belong to a same structural element by means of lengthening of notes and *micropauses*. The *ensemble* rules deal with the case of several voices and force the values attributed to the voices, such as duration and pitch, to be compatible when voices are played together. Other information on automatic music performance can be found at the URL [50].

In the system described in [6], low level and high level rules have been implemented in an *expressiveness* module. An original part of these rules is the *Predictive Amplitude Shaping* algorithm which, according to pitch intervals, computes the global amplitude contour of the note sequence and shapes individually the amplitude of each note. The output of the expressiveness module is called the *MicroScore* and is also the input of the synthesis module It contains all the characteristics necessary for an expressive performance.

7. THE UNIT SELECTION METHOD

The unit selection and concatenation method has been developed for speech synthesis and has led to a large improvement of the quality of the produced voice. It has recently been applied for singing voice synthesis [42], [25], [64]. The principle of the method is to use a large database of recordings of a singing voice segmented into units of one to several successive phones. For a given score with the corresponding lyrics, a sequence of phones aligned on notes is to be produced, along with the specific characteristics of the performance (pitch, duration, timbre, etc.). For each successive part of this sequence, the method searches the database for the *best* unit, i.e. the one which is the closest to the phones and notes to be produced, in the sense of a carefully designed distance.

Even more for singing than for speech, no unit of the database has the exact pitch, nor the exact properties in general, needed for any segment of the sequence. Therefore the *best* unit will have to be modified in pitch, duration, etc., at the risk of degrading the quality. In other respects, if several successive phonemes of the sequence are found in the same order in a unit of the database, this unit is favored since all the *coarticulation* effect are intrinsically taken into account in this unit. But the longer the sequence, the more a large modification is required. In consequence the choice of the unit is a tradeoff between several requirements: to cover the longest sequence of phonemes and to have characteristics (pitch, duration, timbre, etc.) closest to the ones necessary for the given performance. This tradeoff is implemented with a cost function usually including a *concatenation* cost for concatenating units which where not sung in succession (since this can also produce undesirable effects) and a *target* cost for each of the characteristics to be modified in order to attain the characteristics of the performance. Naturally, the global cost is a weighted summation of the individual costs in order to properly balance the different requirements.

How to choose these costs will be explained in the next section. Clearly, the units intrinsically contain the influence of an implicit set of rules applied by the singer with all its training, talent and musical skill. The unit selection and concatenation method is thus a way to replace a large and complicated set of rules by implicit rules from the best performers, and it is often called a *data-driven* concatenative synthesis. But since unit changes always have to be done, the application of another set of rules over the implicit one is a difficulty to overcome in this method.

In order to apply the changes, various analysis-synthesis techniques have been used. The PSOLA technique is simple and robust, as long as changes are limited in range, but does not offer any help for timbre changes. The sinusoidal additive-plus-residual analysis-synthesis technique renders timbre changes easy. But it is more complicated and presents difficulties in the transitions and in the separation of sinusoidal and non-sinusoidal (*residual*) components. Another advantage of the unit selection method is that different voices can easily be produced by means of different databases. One of the drawbacks is the amount of work necessary to build the database and to find the proper cost weights. This is all the more true as a large database improves singing synthesis quality and naturalness [42]. Database construction difficulties will be tackled in the next section.

8. DATABASE CONSTRUCTION

The singing voice has several aspects which differ from speech and have to be taken into account in the unit selection method. Among these aspects, a precise evolution in time of characteristics such as pitch, loudness, etc., within a phone has to be determined (see section 6). But in the case of the unit selection method, these evolutions in time are intrinsically coded in the units and must be included in the selection process. Similarly, the singer's formant is another important characteristic to be taken into account. In consequence, the number of characteristics is increased compared to speech and this makes more difficult the calculation of the optimal cost weights introduced in section 7. Also, in the unit selection method, as mentioned above, the quality of a synthetic singing voice is better with a larger unit database. The large database therefore needed raises difficulties. Recordings with similar acoustic properties should be available, they have to be segmented

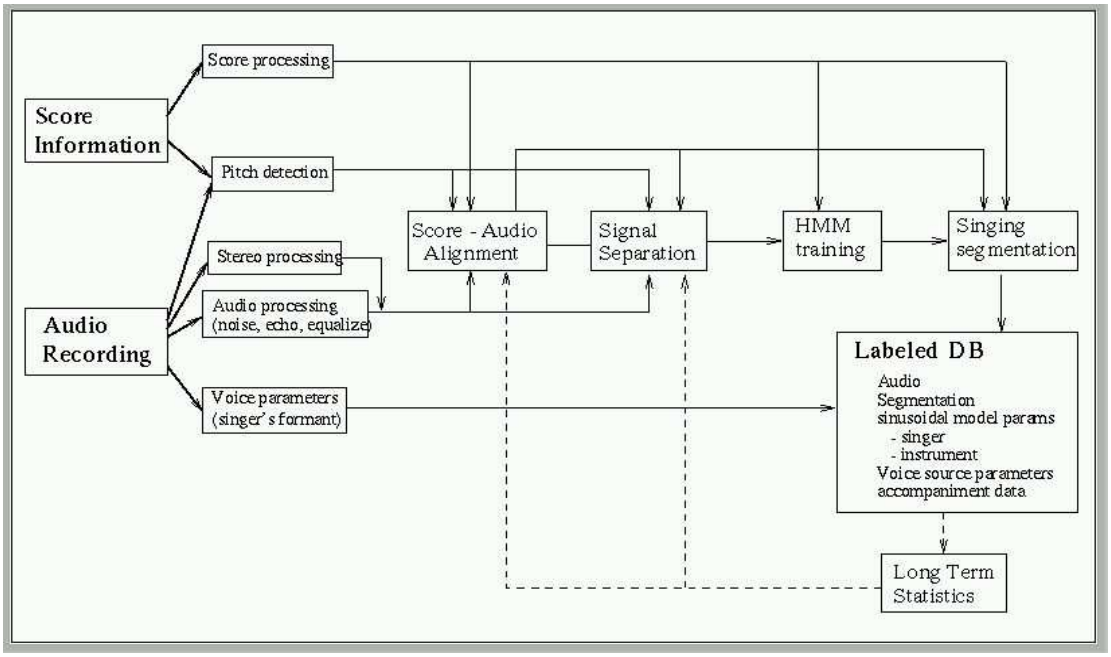

Fig. 4. General scheme of training system ([42])

automatically and optimal cost weights have to be estimated by a learning procedure which can be exaggeratedly time consuming.

In [42], segmentation of singing recordings into units, such as phones and notes, is obtained by use of a Hidden Markov Model (HMM) recognizer designed for speech segmentation. However, the segmentation error rate has been found higher for singing voice than for speech. The segmentation error rate has been lowered by including the score information in the segmentation. More precisely, at first, the recording is time aligned with the score. Then the timing information is used to constrain the segmentation procedure.

Another important achievement in [42] is the improvement of the learning procedure for optimal cost weights, in terms of the quality of the result and of the duration necessary for the procedure to be completed. The training procedure is as follows and applies for a set of original singing voice phrases (Fig. 4). Given an original phrase and one of the synthesized versions possible with the units in the database, a *Cepstral* distance is computed between the original and the synthesis. The idea is to find the cost weights which, for each original, lead to the synthesis sequence of units which has the smallest distance to the original. If the best weight combination is found by an exhaustive search, the total amount of CPU-hours needed is in the hundreds of hours. By a new design of the algorithm [42], this CPU time is reduced by a factor of 500. This allows increasing the search space and the number of original sentences tested, therefore optimizing the cost weights and thus the quality of the synthesis system. Similarly, computation time reductions, by two orders of magnitude, were obtained by modification of the unit selection algorithm itself. Finally, in order to check the different synthesis method being used, listening experiments were conducted in order to evaluate their advantage or drawback.

As said above, the unit selection method requires large recordings of a singer. According to section 2, for some applications the only available recordings are mixtures of the singer 's voice and instruments. In this case it is necessary to separate the voice from the rest, which is an extremely difficult task in general. The task is largely facilitated when the accompaniment is piano only. In [42], a method is developed in order to separate piano accompaniment and singer's voice. This allows the usage, for training and database, of the large number of existing recordings of singers with piano accompaniment.

9. SOUND EXAMPLES

For a better understanding of the state of the art in singing synthesis, sound examples can be heard on web sites of various systems. Some of them are listed here with a brief description of the synthesis system. The system designed at MIT-Media Lab [34], [35] does not start from a score but from an instrumental performance as described in section 2. The parameters (pitch, loudness and brightness) come from the sound of an acoustic instrument such as a violin. A model of a singing voice has been built from recordings by use of the cluster-weighted method [30]. The method used for the voice modification is sinusoidal plus

residual modeling (Cf. section 5). The voice model is played in real time according to the performance parameters, so that the singing voice has the pitch, loudness and brightness envelopes (versus time) issued from the performance. Examples can be heard on:
`http://web.media.mit.edu/~jehan/`
On the opposite, the *Lyricos* system [41] starts from a score and uses rules to compute the expressive contours. For the *articulatory* or *segmental* level, the Unit Selection method is applied. Finally, the ABS/OLA method is used in order to obtain the exact voice characteristics required. Examples can be heard on:
`http://cslu.cse.ogi.edu/tts/research/sing/sing.html`
Similar to Lyricos, the system named *Flinger* (*Festival singer*) starts from score and user-given adjustments. Flinger is based on the *Festival* speech synthesis system [5]. As the previous system does, it uses rules to compute expressive contours. The units employed for synthesis are *diphones* [10]. The synthesis method is also based on sinusoidal modeling. Examples can be heard on:
`http://cslu.cse.ogi.edu/cgi-bin/flinger/show_jukebox.pl?all`
VocalWriter also starts from score and user-given adjustments and applies rules for expressive contours. It offers 85 different human, synthetic and special effects voices. The synthesis method is named RAS for Resonance Articulatory Synthesis. Examples can be heard on:
`http://kaelabs.com/support.htm`
The singing voice system developed at PFU is based on *Spectral Modeling Synthesis* (SMS) and *Excitation plus Resonance* modeling (EpR) [6]. Its input is composed of the melody, lyrics and user controls. Its set of performance and expressivity rules has briefly been presented in section 6. The system also combines a waveform synthesizer (SMS, see section 5.) and a database of recorded voice units. (See section 4).
The singing voice system developed by Y. Meron [42] does not start from a score but from lyrics and performance data, for example extracted from a human performance. Its Unit Selection module has been explained in sections 7 and 8. Waveform synthesis is based on SM-PSOLA quoted in section 5.

10. EVALUATION OF RESULTS

The sound examples listed in the previous section give some idea of the quality achieved by the various available systems. The first important point is that "commercial products are far from real-world application requirements" [6]. These authors also note that their system is more intelligible and natural than VocalWriter but presents a lack of timbre uniformity and is not comparable with a real singer. The system by [42] may be the closest to a real singer, but it should be underlined that it starts not from a score but from a human performance which it portraits and this is known to largely improve the quality of synthetic voice. Meron also lists several possible improvements to his system, among which the use of diphones instead of phones as units in the database and the inclusion of a voice source model in the waveform synthesis method (see section 12). Finally, evaluation is usually done informally. Procedures for systematic and rigorous evaluation do not seem to exist today.

11. CHOIR SINGING SYNTHESIS

There are few studies on choral singing analysis and synthesis among which [75], [76], [77], [78], [3], [4] and [20]. In the opera *K* by composer P. Manoury, a synthetic choir is used, not only to replace a natural one, but also for special effects which could hardly be possible with a natural choir [61]. The synthetic choir is obtained from natural recordings of one single voice for each register. Then, for each register, this voice is duplicated in the number of slightly different *clones* required for the register, typically 6. The clones are obtained by multiple use of a real-time PSOLA algorithm [45], [61] implemented in the Ircam's real-time system *jMax* [15], [36]. Finally each of the so obtained clones are placed at precise locations in space by use of the *Spat*, the Ircam's spatialization system [66]. Each individual clone differs slightly from the others by space location, pitch and timing so as to reproduce the effect of a real group of singers. Other modifications are used, such as voicing alteration, desynchronisation of the voices or movements in space. Some examples of this virtual choir synthesis system can be heard from:
`http://www.ircam.fr/anasyn/Bastille-K/index.html`

12. FURTHER RESEARCH AND DEVELOPMENTS

There are many domains where improvements are foreseen or being studied. Some of the most important will be listed in the following. For better sounding synthesis, the first domain concerns waveform construction. Airflow and acoustics in the glottis are improved in research works such as[47], [48], [49] (examples can be heard in [46]), and an *articulatory* synthesizer proposed in [67], [68] and examples can be heard in [69]. Other aspects which should perhaps be better taken into account are source-filter coupling, non-planar waves and acoustic phenomena in the glottis. For instance, source models have recently received more attention in the works of [31], [32] and [74] and. A pitch synchronous amplitude modulated sinusoidal-residual model has been used in [32]. Another direction of improvement is the analysis of natural recordings in order to obtain better descriptions of the glottal source and of the vocal transfer function. Application of discrete all-pole models has been studied in [31]. Joint estimation of source and filter on the singing voice is obtained in [32]. These authors also use parametric models of the vocal source. By analyzing various types of vocal productions (e.g. normal, pressed, breathy), they have found parameters for these

productions which are then used to drive synthesis models. A very precise estimation of the vocal tract transfer function is proposed in [1]. It is based on the amplitude versus frequency trajectory of singing voice sinusoidal partials induced by vibrato.

The examples quoted in section 9 show that synthetic singing is all the better accepted if the performance has been extracted from a natural performance rather than computed by interpretation rules. This seems to indicate that a lot of work is still to be made in order to improve automatic production of singing specifications from a given musical score. In particular, it is necessary to analyze performances of professionals and learn how they use the voice sound characteristics for expressive purposes.

Similarly, there is a need for labeled and analyzed singing voice databases for research and evaluation, and for their use in unit selection based synthesis software. In this domain, several points need improvements, such as unit selection process, cost function (which should use source characteristics, timbre features, some *quality* criterion), training cost weights and speed of the training process. It should be noticed that progress in source separation would ease learning from existing recordings as mentioned at the end of section 8.

Concerning voice quality, research is being done for example by the Institute of Linguistics at the Utrecht University about the different ways in which to sing a musical phrase, each conveying different nuances [81]. Naturalness [79], realism and expressivity are other points where little research has been done. For instance, emotions in music have been studied in [19] and perception of emotions in singing is investigated in [33]. Finally, most of the research is done on western music while other parts of the world such as India have an old and extremely rich tradition in classical singing which should be studied as well.

13. CONCLUSION

In this paper, the state of the art in singing voice synthesis has been examined. Various points where recent improvements have been obtained are detailed. The most important are probably Waveform synthesizers, unit modification techniques, performance rules, unit selection methods and database construction. Finally a survey has been proposed of some of the domains in which recent important research results have been obtained.

Compared to speech, the singing voice has received much less scientific attention. But the way by which progress has been obtained in the speech domain is worth considering for the singing voice. As an example, evaluation procedures are still to be established and could benefit from experience such as from the [54] at ICAD02 (performance RENdering piano CONcours- workshop for rendered performance). Another aspect which is very important for synthesis of the singing voice is the need for an interdisciplinary approach. As said in the workshop *Music and I.T.*, Feb. 2001 Harrogate, "The development of novel synthesis techniques requires an interaction between cognitive science, musical creativity and engineering".

14. REFERENCES

[1] Arroabarren, I., Amplitude versus frequency trajectory of singing voice sinusoidal partials induced by vibrato, Internal report, Ircam, 2002

[2] Bennett, G., Rodet, X., "Synthesis of the Singing Voice", in *Current Directins in Computer Music Research*, ed. M.V. Mathews & J.R. Pierce, MIT Press, 1989.

[3] Berndtsson, G., Systems for synthesising singing and for enhancing the acoustics of music rooms. Two aspects of shaping musical sounds, doctoral dissertation, KTH, Stockholm 1995.

[4] Berndtsson G., the KTH rule systeme for singing synthesis, Computer Music Journal 20(1):76-91.

[5] Black A.W. and P. Taylor, The Festival Speech Synthesis System: System docmentation, Tec. Rep. HCRC/TR-83, Human Communication Research Center, University of Edinburgh, Scotland, UK, January 1997.

[6] Bonada J., O. Celma, A. Loscos, J. Ortola, X. Serra, Y. Yoshioka, H. Kayama, Y. Hisaminato et H. Kenmochi, Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models, Proc. ICMC 2001, La Habana, Cuba, Sept. 2001

[7] Bresin, R. (2000) Virtual Virtuosity - Studies in automatic music performance. Doctoral Dissertation, Speech Music and Hearing, Stockholm: KTH, TRITA-TMH 2000:9, ISSN 1104-5787,ISBN 91-7170-643-7

[8] `http://www.speech.kth.se/music/publications/thesisrb/`

[9] Bresin, R., & Friberg, A. (2000). Emotional coloring of computer-controlled music performance. Computer Music Journal, 24, 44-62.

[10] Charpentier F. and E. Moulines, Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones, Speech Communication Vol. 9., pp 453-467, Dec. 1990.

[11] Ciocea, S. and J. Schoentgen, Semi-analytic formant-to-area mapping, Etudes et Travaux — ILVP/ULB, Ao t 1998.

[12] P. R. Cook, Singing voice synthesis: History, current work, and future directions, Computer Music Journal, vol. 20, no. 3, Fall 1996.

[13] Toward the Perfect Audio Morph, First European COST conference on Digital Audio Effects, Barcelona, 1998.

[14] P. Depalle, G. Garc & X. Rodet, A virtual Castrato (!?), Proc. Int. Computer Music Conference, Aarhus, Denmark, Oct. 1994.

[15] D chelle,F. et al. (1999) , jMax: an environnement for real-time musical applications. Computer Music Journal 23(3), 50-58.

[16] Desain, P.W.M., & Honing, H.J. (1997). How to evaluate generative models of expression in music performance. In K.Hirata (Ed.), Issues in Ai and music evaluation and assessment. Workshop notes. International joint conference on artificial intellingence (pp. 5-7). Nagaya, Japan.

[17] Desain P., Vibrato and portamento, hypotheses and tests. (1999). Acustica 348. ISSN:1436-7947.

[18] Expressing emotion in music
http://www.speech.kth.se/music/performance/performance_emotion.html
[19] http://www.speech.kth.se/~roberto/emotion/
[20] Fagnan, L., The acoustical effects of the core principles of the bel canto method on choral singing.
[21] http://www.ircam.fr/anasyn/PRESENTATIONS/FAGNAN/index-e.html
[22] G. Fant, J. Liljencrants and Q. Lin . A four parameter model of glottal flow. STL-QPSR , 4:1-13, 1985
[23] G. Fant, The LF-model revisited. Transformations and frequency domain anals ysis
[24] STL-QPSR , 2-3:119-156, 1995
[25] http://cslu.cse.ogi.edu/cgi-bin/flinger/show_jukebox.pl?all
[26] Friberg, A., Sundberg, J. & Fryd n,L. (1994). Rules for musical performance, in Information Technology and Music, CD-ROM produced by the Royal Swedish Academy of Engineering Science. 1994
[27] Friberg, A., Schoonderwaldt, E., Juslin, P. N., & Bresin, R. (2002). Automatic real-time extraction of musical expression. Manuscript submitted for publication.
[28] Friberg, A. and Sundstr m,A. (2002). "Swing ratios and ensemble timing in jazz performance: Evidence for a common rhythmic pattern". Music Perception 19(3), 333-349.
[29] George E. B. and M. J. T. Smith, Speech analysis/synthesis and modification using an analysis-by-synthesis/overlapp-add model, IEEE Trans. on Speech and Audio Processing, Vol. 5, pp. 389-406, Sep. 1997.
[30] Gershenfeld N.A., B. Schoner and E. M tois, Cluster-weighted modeling for time series analysis. Nature, 37:329-332, (1999).
[31] Heinrich, N., Etude de la source glottique en voix parl et chant e, Th se de l Universit Paris 6, Paris, 30 Novembre 2001.
[32] Hui-Ling L., Toward a hig hquality singing synthesizer with vocal texture control, PhD Thesis, Stanford Univ., 2002.
[33] Jansens S., G. Bloothooft and G. de Krom, Perception and acoustics of emotions in singing, Proc 5th Eurospeech, Rhodes, 1997, IV:2155-218.
[34] Jehan T., B. Schoner, An Audio-Driven Perceptually Meaningful Timbre Synthesizer, Proceedings International Computer Music Conference. Havana, Cuba, 2001
[35] Jehan T., B. Schoner, An Audio-Driven, Spectral Analysis-Based, Perceptual Synthesis Engine, Audio Engineering Society, Proceedings of the 110th Convention. Amsterdam, The Netherlands, 2001
[36] http://www.ircam.fr/equipes/temps-reel/jmax/en/index.php3
[37] Juslin, P. N.,Feedback learning in musical expression http://www.psyk.uu.se/forskning/projekt.html#pj
[38] Juslin, P. N., Friberg, A., & Bresin, R. (2002). Toward a computational model of expression in performance: The GERM model. Musicae Scientiae specal issue 2001-2002, 63-122.
[39] Y. E. Kim, Structured Encoding of the Singing Voice using Prior Knowledge of the Musical Score, Proc. 1999 IEEE Workshop on Application of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999.
[40] Lee M. E. and M. J. T. Smith, Digital Voice Synthesis using a new alternating reflection Model, IEEE-ISCAS 2002.
[41] Macon M.W., L. Jensen-Link, J. Oliviero, M.A. Clements and E. Bryan George, Concatenation based MIDI-to-Singing Voice Synthesis, 103rd Meeting of the AES, Sep. 1997 - 1ES Preprint 4591.
[42] Meron Y., High quality singing synthesis using the selection-base synthesis scheme, PhD dissertation, Univ. of Tokyo, 1999.
[43] Report of the EPSRC Music and Information Technology workshop, White Hart Conference Center, Harrogate, 21-22 February 2001
[44] Peeters, G. (1998). Analyse-Synth sedes sons musicaux par la m thodePSOLA. In proc. Journ es Informatique Musicale, Agelonde, France.
[45] G. Peeters, X. Rodet, SINOLA : A New Method for Analysis/Synthesis using Spectrum Distorsion, Phase and Reassigned Spectrum, ICMC: International Computer Music Conference, (Pekin, Chine, Octobre 1999).
[46] http://www.icp.inpg.fr/~pelorson/sons.html
[47] Pelorson X., Hirschberg A., Wijnands A.P.J., Bailliet H., Vescovi C., Castelli E. (1996) Description of the flow through the vocal cords during phonation. Application to voiced sounds synthesis.Acta Acustica, 82, 358-361.
[48] Pelorson X., Hofmans G.C.J., Ranucci M., Bosch R.C.M. (1997) On the fluid mechanics of bilabial plosives.Speech Communication, 22, 155-172.
[49] Pelorson X., Msallam R., Gilbert J, Hirschberg A. " Fluid dynamic aspects of human voice and brass instruments: implications for sound synthesis", in Proc. ICA/ASA, Seattle, 1998.
[50] Rules for Music Performance
http://www.speech.kth.se/music/performance/
[51] Prame E. Measurements of the vibrato rate of ten singers. J. Acoust. Soc. Am. 96 (1994), 1979-1984.
[52] Prame E. Vibrato extent and intonation in professional Western lyric singing. J. Acoust. Soc. Am. 102 (1997), 616-621.
[53] Synthesis of the Aria The Queen of the Night from Mozart's Opera The Magic Flute
http://www.ircam.fr/anasyn/reine.html
[54] Performance RENdering piano CONcours- workshop for rendered performance, ICAD02, ATR, Kyoto, July 6, 2002.
[55] Rodet, X., Musical Sound Signals Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models, Applied Signal Processing (1997) 4:131-141
[56] Rodet, X., "Time Domain Formant-Wave- Function Synthesis", Cambridge, Massachusets, Computer Music Journal, Vol 8, n¡3, 1984.
[57] Rodet, X., Yves Potard, Jean-Baptiste Barriere, ''The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General,'' Computer Music Journal, Vol. 8(3), pp. 15-31, 1984.
[58] X. Rodet, A. Lef vre: The Diphone program New features, new synthesis methods and experience of musical use , proc. Int. Comp. Music Conference, Thessaloniki, 1997.
[59] Rodet X. and D. Schwarz, Spectral envelopes and additive+residual analysis-synthesis, to appear in J. Beauchamp ed. The Sound of Music. Springer N.Y. to be published

[60] Ross J, J. Sundberg, Syllable and tone boundaries in singing, 4th Pan European Voice Conference held in the beautiful city of Stockholm on August 23-26, 2001

[61] Norbert Schnell, Geoffroy Peeters, Serge Lemouton, Philippe Manoury, Xavier Rodet: Synthesizing a choir in real-time using Pitch-Synchronous Overlap Add (PSOLA) ICMC2000

[62] Schwarz, D., Rodet, X., Spectral envelope estimation, representation, and morphing for sound analysis, transformation, and synthesis. Proc. Int. Comp. Music Conf., ICMC99, Beijing, Oct. 99, pp. 351-354.

[63] http://www.ircam.fr/anasyn/DOCUMENTATIONS/specenv/

[64] Schwarz, D. ¨°A System for Data-Driven Concatenative Sound Synthesis°¨. Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx-00), Verona, Italy, December 7-9, 2000.

[65] Serra, X., A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. Philosophy Dissertation, Stanford University, Oct. 1989

[66] http://www.ircam.fr/produits/logiciels/spat.

[67] Story, B.H., and Titze, I.R., 1998. Parameterization of vocal tract area functions by empirical orthogonal modes, J. Phonetics, (26(3), 223-260.

[68] Story, B.H., Titze, I.R., and Long, R., "Simulation of sentence-level speech based on measured area functions", Proc. of the ICA/ASA, Seattle, WA, June, 1998, 2663-2664.

[69] Singing simulated by B. Story, I. Titze, and E. Hunter. http://web1.dcpa.org/brad_html/bstory.html

[70] Sundberg, J., The science of the singing voice, Northern Illinois University Press, Dekalb, Illinois, 1987.

[71] Grammars for music performance, Proc. of a KTH symposium May 1995, Stockholm, KTH, Dept. of Speech Comm. and Music Acoust., J. Sundberg & A. Friberg eds..

[72] Sundberg 96] Sundberg J, Prame E, Iwarsson J. Replicability and accuracy of pitch patterns in professionalsingers. Chapter 20 in P Davis & N Fletcher, ed:s, Vocal Fold Physiology, Controlling Complexity and Chaos, San Diego: Singular Publ Group 1996, 291-306.

[73] Sundberg J, Keynote - Sounds of singing. A matter of mode, style, accuracy, and beauty, 4th Pan European Voice Conference held in the beautiful city of Stockholm on August 23-26, 2001

[74] Ternstr m, S., Analysis and Simulation of the Glottal Oscillator, http://www.speech.kth.se/music/staff/sten_projects.html

[75] Ternstr m, S &J. Sundberg. Formant frequencies in choir singers. Journal of Acoustical Society of America 86 (2), 517-522 (1989).

[76] Ternstr m,S., Long-time average spectrum characteristics of different choirs in different rooms. Voice (UK), 2, 55-77 (1993).

[77] Ternstr m, S., Preferred self-to-other ratios in choir singing. J Acoust Soc Am, June 1999, 105(6), 3563-3574.

[78] Ternstr mS, Karna DR. Acoustics of Choir Singing. 5th International Congress of VoicTeachers, Helsinki, August 2001.

[79] Ternstr m, S, Session on Naturalness in Synthesized Speech and Music, 143rd ASA meeting, Pittsburgh, June 3-7, 2002.

[80] Timmers, R. and P. Desain, Vibrato°: questions and anspects from musicians and science, Proc. Of the sixth ICMPC, Keele, 2000.

[81] Institute of Linguistics at the Utrecht University http://www-uilots.let.uu.nl/