Expression Control in Singing Voice Synthesis

Features, Approaches, Evaluation, and Challenges

Martí Umbert, Jordi Bonada, Masataka Goto, Tomoyasu Nakano, and Johan Sundberg

M. Umbert and J. Bonada are with the Music Technology Group (MTG) of the Department of Information and Communication Technologies, Universitat Pompeu Fabra, 08018 Barcelona, Spain, e-mail: {marti.umbert, jordi.bonada}@upf.edu

M. Goto and T. Nakano are with the Media Interaction Group, Information Technology Research Institute (ITRI) at the National Institute of Advanced Industrial Science and Technology (AIST), Japan, e-mail: {m.goto, t.nakano}@aist.go.jp

J. Sundberg is with the Voice Research Group of the Department of Speech, Music and Hearing (TMH) at the Royal Institute of Technology (KTH), Stockholm, Sweden, e-mail: jsu@csc.kth.se

In the context of singing voice synthesis, expression control manipulates a set of voice features related to a particular emotion, style, or singer. Also known as performance modeling, it has been approached from different perspectives and for different purposes, and different projects have shown a wide extent of applicability. The aim of this article is to provide an overview of approaches to expression control in singing voice synthesis. Section I introduces some musical applications that use singing voice synthesis techniques to justify the need for an accurate control of expression. Then, expression is defined and related to speech and instrument performance modeling. Next, Section II presents the commonly studied set of voice parameters that can change perceptual aspects of synthesized voices. Section III provides, as the main topic of this review, an up-to-date classification, comparison, and description of a selection of approaches to expression control. Then, Section IV describes how these approaches are currently evaluated and discusses the benefits of building a common evaluation framework and adopting perceptually-motivated objective measures. Finally, Section V discusses the challenges that we currently foresee.

Table 1: Research projects using singing voice synthesis technologies.

Project	Website
Cantor	http://www.virsyn.de
Cantor Digitalis	http://www.cantordigitalis.limsi.fr
Flinger	http://www.cslu.ogi.edu/tts/flinger
Lyricos	http://www.cslu.ogi.edu/tts/demos
Orpheus	http://www.orpheus-music.org/v3
Sinsy	http://www.sinsy.jp
Symphonic Choirs Virtual Instrument	http://www.soundsonline.com/Symphonic-Choirs
VocaListener	https://staff.aist.go.jp/t.nakano/VocaListener
VocaListener (product version)	http://www.vocaloid.com/lineup/vocalis
VocaListener2	https://staff.aist.go.jp/t.nakano/VocaListener2
Vocaloid	http://www.vocaloid.com
VocaRefiner	https://staff.aist.go.jp/t.nakano/VocaRefiner
VocaWatcher	https://staff.aist.go.jp/t.nakano/VocaWatcher

I. Introduction

In this section we put into context the expression control in singing voice synthesis. First, we describe the main building blocks of these technologies. Then, we define expression in music performance and singing. Finally, we give an insight into how this area of research relates to the study of expression in the speech and instrumental music performance modeling communities.

A. Singing voice synthesis systems

During recent decades, several applications have shown how singing voice synthesis technologies can be of interest for composers [1] [2]. Technologies for the manipulation of voice features have been increasingly used to enhance tools for music creation and post-processing, singing live performance, to imitate a singer, and even to generate voices difficult to produce naturally (e.g. castrati). More examples can be found with pedagogical purposes or as tools to identify perceptually relevant voice properties [3]. These applications of the so-called music information research field may have a great impact on the way we interact with music [4]. Examples of research projects using singing voice synthesis technologies are listed in Table 1.

The generic framework of these systems is represented in Fig. 1, based on [5]. The input may consist of the score (e.g. note sequence, contextual marks related to loudness, or note transitions), lyrics, and the intention (e.g. the style or emotion). Intention may be derived from the lyrics and

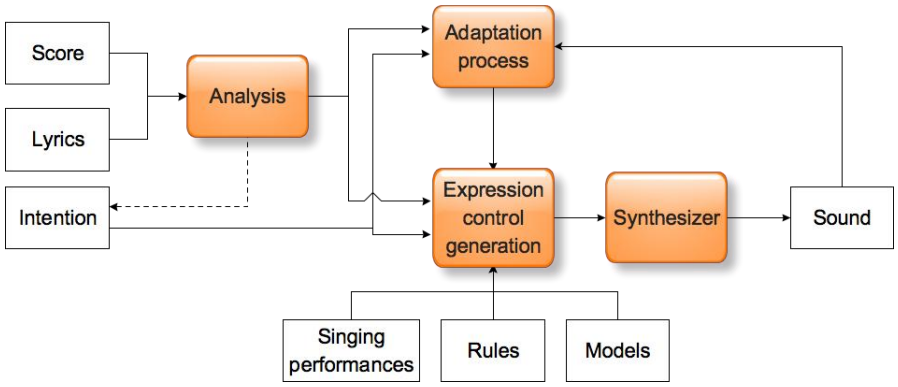


Fig. 1: Generic framework blocks for expression control.

score content (dashed line). The input may be analyzed to get the phonetic transcription, the alignment with a reference performance, or contextual data. The expression control generation block represents the implicit or explicit knowledge of the system as either a set of reference singing performances, a set of rules, or statistical models. Its output is used by the synthesizer to generate the sound, which may be used iteratively to improve the expression controls.

A key element of such technologies is the singer voice model [1] [2] [6], although it is out of the scope of this publication to describe it in depth. For the purpose of this article, it is more interesting to classify singing synthesis systems with respect to the control parameters. As shown in Table 2, those systems are classified into model-based and concatenative synthesizers. While in signal models the control parameters are mostly related to a perception perspective, in physical models these are related to physical aspects of the vocal organs. In concatenative synthesis, a cost criterion is used to retrieve sound segments (called units) from a corpus which are then transformed and concatenated to generate the output utterance. Units may cover a fixed number of linguistic units, e.g. diphones that cover the transition between two phonemes, or a more flexible and wider scope. In this case, control parameters are also related to perceptual aspects.

Within the scope of this review, we focus on the perceptual aspects of the control parameters which are used to synthesize expressive performances by taking a musical score, lyrics or an optional human performance as the input. This review therefore, does not discuss voice conversion and morphing in which input voice recordings are analyzed and transformed [7] [8].

Table 2: Singing voice synthesis systems and control parameters.

	Singing synthesis systems			
	Model-based synthesis		Concatenative synthesis	
	Signal models	Physical models	Fixed length units	Non uniform length units
Parameters	F0, resonances (centre frequency and bandwidth), sinusoid frequency, phase, amplitude, glottal pulse spectral shape, and phonetic timing	Vocal apparatus related parameters (tongue, jaw, vocal tract length, and tension, sub-glottal air pressure, phonetic timing)	F0, amplitude, timbre, and phonetic timing	

B. Expression in musical performance and singing

Expression is an intuitive aspect of a music performance, but complex to define. In [5], it is viewed as "*the strategies and changes which are not marked in a score but which performers apply to the music*" (p. 2). In [9], expression is "*the added value of a performance and is part of the reason that music is interesting to listen to and sounds alive*" (p. 1). A quite complete definition is given in [10], relating the liveliness of a score to "*the artist's understanding of the structure and 'meaning' of a piece of music, and his/her (conscious or unconscious) expression of this understanding via expressive performance*" (p. 150). From a psychological perspective, Juslin [11] defines it as "*a set of perceptual qualities that reflect psychophysical relationships between 'objective' properties of the music, and 'subjective' impressions of the listener*" (p. 276).

Expression has a key impact on the perceived quality and naturalness. As pointed out by Ternström [13], "*even a single sine wave can be expressive to some degree if it is expertly controlled in amplitude and frequency*". Ternström says that musicians care more about instruments being adequately expressive than sounding natural. For instance, in Clara Rockmore's performance of Vocalise by Sergei Vasilyevich Rachmaninoff a skillfully controlled Theremin expresses her intentions to a high degree[1], despite the limited degrees of freedom.

In the case of the singing voice, achieving a realistic sound synthesis implies controlling a wider set of parameters than just amplitude and frequency. These parameters can be used by a singing voice synthesizer or to transform a recording. From a psychological perspective, pitch

[1] All cited sounds have been collected in: www.mtg.upf.edu/publications/ExpressionControlinSingingVoiceSynthesis

contour, vibrato features, intensity contour, tremolo, phonetic timing, and others related to timbre are the main control parameters that are typically used to transmit a message with a certain mood or emotion [12] and shaped by a musical style [14]. These are described in detail in Section II.

Nominal values for certain parameters can be inferred from the musical score, such as note pitch, dynamics and note duration and its articulation like staccato or legato marks. However, these values are not intrinsically expressive per se. In other words, expression contributes to the differences between these values and a real performance. Different strategies for generating expression controls are explained in Section III.

It is important to note that there is more than one acceptable expressive performance for a given song [1] [3] [15]. Such variability complicates the evaluation and comparison of different expression control approaches. This issue is tackled in Section IV. Besides singing, expression has been studied in speech and instrumental music performance, as presented in the next section.

C. Connection to speech and instrumental musical performance

There are several common aspects in performing expressively through singing voice, speech, and musical instruments. In speech, the five acoustic attributes of prosody have been widely studied [16], for instance to convey emotions [17]. The most studied attribute is the fundamental frequency (F0) of the voice source signal. Timing is the acoustic cue of rhythm and it is a rather complex attribute given the number of acoustic features it is related to [16] (p. 43). Other attributes are intensity, voice quality (related to the glottal excitation), and articulation (largely determined by the phonetic context and speech rate).

Expressive music performance with instruments has also been widely studied. Several computational models are reviewed in [18], like the KTH model, which is based "*on performance rules that predict the timing, dynamics, and articulation from local musical context*" (p. 205). The Todd model links the musical structure to a performance with simple rules like measurements of human performances. The Mazzola model analyzes musical structure features like tempo and

melody and iteratively modifies expressive parameters of a synthesized performance. Finally, a machine-learning model discovers patterns within large amounts of data, it focuses for instance on timing, dynamics, and more abstract structures like phrases, and manipulates them via tempo, dynamics, and articulation. In [5], 30 more systems are classified into non-learning methods, linear regression, artificial neural networks, rule/case-based learning models among others.

In this review, we adopt a signal processing perspective to focus on the acoustic cues that convey a certain emotion or evoke a singing style in singing performances. As mentioned in [12], "*vocal expression is the model on which musical expression is based*" (p. 799), which highlights the topic relevance for both the speech and the music performance community. Since there is room for improvement, the challenges that we foresee are described in Section V.

II. Singing voice performance features

In Section I.B we introduced a wide set of low-level parameters for singing voice expression. In this section we relate them to other musical elements. Then, the control parameters are described, and finally, we illustrate them by analyzing a singing voice excerpt.

A. Feature classification

As in speech prosody, music can also be decomposed into various musical elements. The main musical elements such as melody, dynamics, rhythm, and timbre are built upon low-level acoustic features. The relationships between these elements and the acoustic features can be represented in several ways [19] (p. 44). Based on this, Table 3 relates the commonly modeled acoustic features of singing voice to the elements to which they belong. Some acoustic features spread transversally over several elements. Some features are instantaneous such as F0 and intensity frame values, some span over a local time window like articulation and attack, and others have a more global temporal scope like F0 and intensity contours, or vibrato and tremolo features.

Next, for each of these four musical elements, we provide introductory definitions to their acoustic features. Finally, these are related to the analysis of a real singing voice performance.

Table 3: Classification of singing voice expression features.

Melody	Dynamics	Rhythm	Timbre
Vibrato and tremolo (depth and rate)		Pauses	Voice source
Attack and release		Phoneme time-lag	Singer's formant
Articulation	Phrasing		Sub-harmonics
F0 contour	Intensity contour	Note/phoneme onset/duration	Formant tuning
F0 frame value	Intensity frame value	Timing deviation	Aperiodicity spectrum
Detuning		Tempo	

B. Melody related features

The F0 contour, or the singer's rendition of the melody (note sequence in a score), is the sequence of F0 frame-based values [20]. F0 represents the "*rate at which the vocal folds open and close across the glottis*", and acoustically it is defined as *"the lowest periodic cycle component of the acoustic waveform*" [12] (p. 790). Perceptually it relates to pitch, defined as "*the aspect of auditory sensation whose variation is associated with musical melodies*" [21] (p. 2). In the literature, however, pitch and F0 terms are often used indistinctly to refer to F0.

The F0 contour is affected by micro-prosody [22], that is to say, fluctuations in pitch and dynamics due to phonetics (not attributable to expression). While certain phonemes like vowels may have stable contours, other phonemes such as velar consonants may fluctuate due to articulatory effects.

A skilled singer can show the expressive ability through the melody rendition and modify it more expressively than unskilled singers. Pitch deviations from the theoretical note can be intentional as an expressive resource [3]. Moreover, different articulations, that is to say the F0 contour in a transition between consecutive notes, can be used expressively. For example, in 'staccato' short pauses are introduced between notes. In Section F the use of vibratos is detailed.

C. Dynamics related features

As summarized in [12], intensity (related to the perceived loudness of the voice) is a "*measure of energy in the acoustic signal*" usually from the waveform amplitude (p. 790). It "*reflects the effort required to produce the speech*" or singing voice, and is measured by energy at a frame

level. A sequence of intensity values provides the intensity contour, correlated to the waveform envelope and the F0 since energy increases with the F0 so to produce a similar auditory loudness [23]. Acoustically, vocal effort is primarily related to the spectrum slope of the glottal sound source rather than to the overall sound level. Tremolo may also be used, as detailed in Section F.

Micro-prosody has also an influence on intensity. The phonetic content of speech may produce intensity increases as in plosives or reductions like some unvoiced sounds.

D. Rhythm related features

Perception of rhythm involves cognitive processes such as "*movement, regularity, grouping, and yet accentuation and differentiation*" [24] (p. 588), where it is defined as "*the grouping and strong/weak relationships*" amongst the beats, or "*the sequence of equally spaced phenomenal impulses which define a tempo for the music*". Tempo corresponds to the number of beats per minute. In real life performances, there are timing deviations from the nominal score [12].

Similarly to the role of speech rate in prosody, phoneme onsets are also affected by singing voice rhythm. Notes and lyrics are aligned so that the first vowel onset in a syllable is synchronized with the note onset and any preceding phoneme in the syllable is advanced [3] [25].

E. Timbre related features

Timbre depends mainly on the vocal tract dimensions and on the mechanical characteristics of the vocal folds which affect the voice source signal [23]. Timbre is typically characterized by an amplitude spectrum representation, and often decomposed into source and vocal tract components.

The voice source can be described in terms of its F0, amplitude, and spectrum (vocal loudness and mode of phonation). In the frequency domain, the spectrum of the voice source is generally approximated by an average slope of -12 dB/octave, but typically varies with vocal loudness [23]. Voice source is relevant for expression and used differently among singing styles [14].

The vocal tract filters the voice source emphasizing certain frequency regions or formants.

Although formants are affected by all vocal tract elements, some have a higher effect on certain formants. For instance, the first two formants are related to the produced vowel, with the first formant being primarily related to the jaw opening and the second formant to the tongue body shape. The next three formants are rather related to timbre and voice identity, with the third formant being particularly influenced by the region under the tip of the tongue and the fourth to the vocal tract length and dimensions of the larynx [23]. In western male operatic voices the 3rd, 4th, and 5th typically cluster, producing a marked spectrum envelope peak around 3 kHz, the so-called singer's formant cluster [23]. This makes it easier to hear the singing voice over a loud orchestra. The affected harmonic frequencies (multiples of F0) are radiated most efficiently towards the direction where the singer is facing, normally the audience.

Changing modal voice into other voice qualities can be used expressively [26]. Rough voice results from a random modulation of the F0 of the source signal (jitter) or of its amplitude (shimmer). In growl voice sub-harmonics emerge due to half periodic vibrations of the vocal folds and in breathy voices the glottis does not completely close, increasing the presence of aperiodic energy.

F. Transverse features

Several features from Table 3 can be considered transversal given that they spread over several elements. In this section we highlight the most relevant ones.

Vibrato is defined [23] as a nearly sinusoidal fluctuation of F0. In operatic singing, it is characterized by a rate that tends to range from 5.5 to 7.5 Hz and a depth around ± 0.5 or 1 semitones. Tremolo [23] is the vibrato counterpart observed in intensity. It is caused by the vibrato oscillation when the harmonic with the greatest amplitude moves in frequency, increasing and decreasing the distance to a formant, thus making the signal amplitude vary. Vibrato may be used for two reasons [23] (p. 172). Acoustically, it prevents harmonics from different voices from falling into close regions and producing beatings. Also, vibratos are difficult to produce under

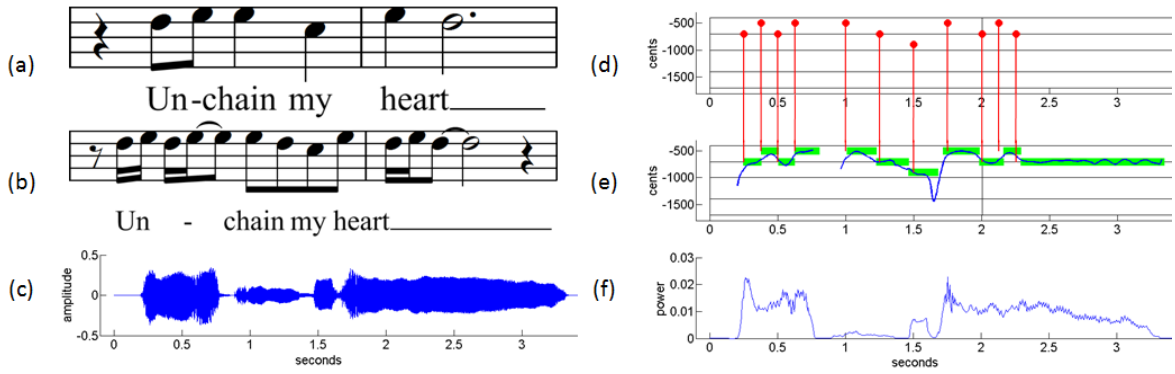


Fig. 2: Expression analysis of a singing voice sample (a) score, (b) modified score, (c) waveform, (d) note onsets and pitch, (e) extracted pitch and labeled notes, (f) extracted energy.

phonatory difficulties like pressed phonation. Aesthetically, vibrato shows that the singer is not running into such problems when performing a difficult note or phrase like high pitched notes.

Attack is the musical term to describe the pitch and intensity contour shapes and duration at the beginning of a musical note or phrase. Release is the counterpart of attack, referring to the pitch and intensity contour shapes at the end of a note or phrase.

As summarized in [27], grouping is one of the mental structures that are built while listening to a piece that describes the hierarchical relationships between different units. Notes, the lowest-level unit, are grouped into motifs, motifs into phrases, and phrases into sections. The piece is the highest-level unit. Phrasing is a transversal aspect that can be represented as an "*arch-like shape*" applied to both tempo and intensity during a phrase [15] (p. 149). For example, a singer may increase tempo at the beginning of a phrase or decrease it at the end for classical music.

G. Singing voice performance analysis

To illustrate the contribution of the acoustic features to expression, we analyze a short excerpt[2] of a real singing performance. It contains clear expressive features like vibrato in pitch, dynamics, timing deviations in rhythm, and growl in timbre. The result of the analysis is shown in Fig. 2 and Fig. 3 (dashed lines indicate harmonic frequencies and the circle is placed at sub-harmonics). The original score and lyrics are shown in Fig. 2a, where each syllable corresponds to one note except

[2] Excerpt from "*Unchain my heart*" song: www.mtg.upf.edu/publications/ExpressionControlinSingingVoiceSynthesis

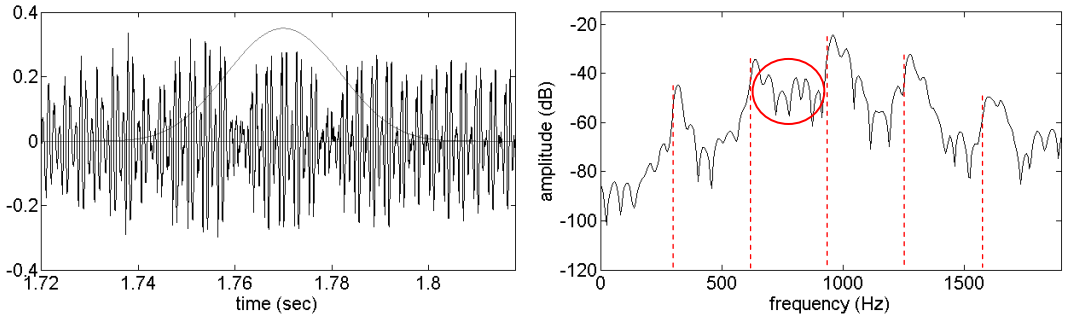


Fig. 3: Growl analysis of a singing voice sample: (a) waveform and (b) spectrum

the first and last ones, which correspond to two notes. The singer introduces some changes like ornamentation and syncopation, represented in Fig. 2b. In Fig. 2c the note pitch is specified by the expected frequency in cents and the note onsets are placed at the expected time using the note figures and a 120 bpm tempo. Fig. 2d shows the extracted F0 contour in blue and the notes in green. The micro-prosody effects can be observed, for example in a pitch valley during the attack to the *'heart'* word. At the end, vibrato is observed. The pitch stays at the target pitch for a short period of time, especially in the ornamentation notes.

In a real performance, tempo is not generally constant throughout a score interpretation. In general, beats are not equally spaced through time, leading to tempo fluctuation. Consequently, note onsets and rests are not placed where expected with respect to the score. In Fig. 2d, time deviations can be observed between the labeled notes and the projection colored in red from the score. Also, note durations differ from the score.

The recording's waveform and energy, aligned to the estimated F0 contour, are drawn in Fig. 2e and in Fig. 2f, respectively. The intensity contour increases/decays at the beginning/end of each segment or note sequence. Energy peaks are especially prominent at the beginning of each segment, since a growl voice is used and increased intensity is needed to initiate this effect.

We can take a closer look at the waveform and spectrum of a windowed frame, as in Fig. 3. In the former, we can see the pattern of a modulation in amplitude or macro-period which spans over several periods. In the latter we can see that, for the windowed frame, apart from the

frequency components related to F0 around 320 Hz, five sub-harmonic components appear between F0 harmonics, which give the "growl" voice quality. Harmonics are marked with a dashed line and sub-harmonics between the second and the third harmonics with a red circle.

If this set of acoustic features is synthesized appropriately, the same perceptual aspects can be decoded. Several approaches that generate these features are presented in the next section.

III. Expression control approaches

In Section II, we defined the voice acoustic features and related them to aspects of music perception. In this section we focus on how different approaches generate expression controls. First, we propose a classification of the reviewed approaches and next we compare and describe them. As it will be seen, acoustic features generally map one-to-one to expressive controls at the different temporal scopes, and the synthesizer is finally controlled by the lowest-level acoustic features (F0, intensity, and spectral envelope representation).

A. Classification of approaches

In order to see the big picture of the reviewed works on expression control, we propose a classification in Fig. 4. Performance-driven approaches use real performances as the control for a synthesizer, taking advantage of the implicit rules that the singer has applied to interpret a score. Expression controls are estimated and applied directly to the synthesizer. Rule-based methods derive a set of rules that reflect the singers' cognitive process. In analysis-by-synthesis, rules are evaluated by synthesizing singing voice performances. Corpus-derived rule-based approaches generate expression controls from the observation of singing voice contours and imitating their behavior. Statistical approaches generate singing voice expression features using techniques such as Hidden Markov Models (HMMs). Finally, unit selection-based approaches select, transform, and concatenate expression contours from excerpts of a singing voice database. Approaches using a training database of expressive singing have been labeled as corpus-based methods.

The difficulties of the topic reviewed in this article center on how to generate control

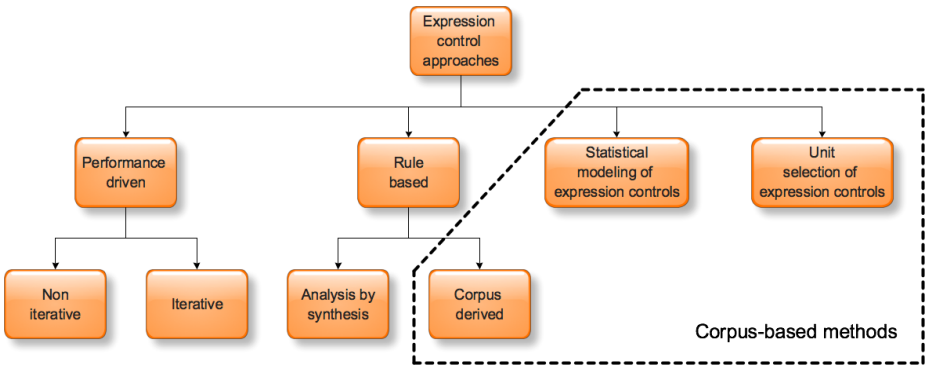


Fig. 4: Classification of Expression Control Methods in Singing Voice Synthesis.

parameters which are perceived as natural. The success of conveying natural expression depends on a comprehensive control of the acoustic features introduced in Section II. Currently, statistical approaches are the only type of system that jointly model all the expression features.

B. Comparison of approaches

In this article we review a set of works which model the features that control singing voice synthesis expression. Physical modeling perspective approaches can be found for instance in [28].

Within each type of approach in Fig 4, there are one or more methods for expression control. In Table 4 we provide a set of items we think can be useful for comparison. From left to right, *Type* refers to the type of expression control from Fig. 4 to which the *Reference* belongs. In *Control features* we list the set of features that the approach deals with. Next, we provide the type of *Synthesizer* used to generate the singing voice, followed by the emotion, style or sound to which the expression is targeted. Also, we detail the *Input* to the system (score, lyrics, tempo, audio recording, etc). The last column lists the *language* dependency of each method, if any.

We have collected[3] samples from most of the approaches in order to help to easily listen to the results of the reviewed expression control approaches. The reader will observe several differences among them. First, some samples consist of a cappella singing voice, and others are presented with background music which may mask the synthesized voice and complicate the perception of the generated expression. Second, samples correspond to different songs, which makes it difficult

[3] www.mtg.upf.edu/publications/ExpressionControlinSingingVoiceSynthesis

to compare approaches. Concerning the lyrics, though in most cases these belong to a particular language, in some the lyrics are made by repeating the same syllable, such as /la/. We believe that the evaluation of a synthesized song can be performed more effectively in a language spoken by the listener. Finally, the quality of the synthetic voice is also affected by the type of synthesizer used in each sample. The difficulties in comparing them and the subsequent criticism are discussed in the evaluation and challenges sections.

C. Performance-driven approaches

These approaches use a real performance to control the synthesizer. The knowledge applied by the singer, implicit in the extracted data, can be used in two ways. In the first one, control parameters like F0, intensity, timing, etc from the reference recording are mapped to the input controls of the synthesizer so that the rendered performance follows the input signal expression. Alternatively, speech audio containing the target lyrics is transformed in order to match pitch and timing of the input score. Fig. 5 summarizes the commonalities of these approaches on the inputs (reference audio, lyrics, and possibly the note sequence) and intermediate steps (phonetic alignment, acoustic feature extraction, and mapping) that generate internal data like timing information, acoustic features, and synthesizer controls used by the synthesizer.

In Table 5 we summarize the correspondence between the extracted acoustic features and the synthesis parameters for each of these works. The extracted F0 can be mapped directly into the F0 control parameter, processed into a smoothed and continuous version, or split into the MIDI note, pitch bend, and its sensitivity parameters. Vibrato can be implicitly modeled in the pitch contour, extracted from the input, or selected from a database. Energy is generally mapped directly into dynamics. From the phonetic alignment, note onsets and durations are derived, mapped directly to phoneme timing, or mapped either to onsets of vowels or voiced phonemes. Concerning timbre, some approaches focus on the singer's formant cluster and in a more complex case the output timbre comes from a mixture of different voice quality databases.

Table 4: Comparison of approaches for Expression control in Singing Voice Synthesis.

Type	Reference	Control features	Synthesizer	Style or emotion	Input	Language
Performance-driven	Y. Meron (1999) [29]	Timing, F0, intensity, singer's formant cluster	Unit-selection	Opera	Score, singing voice	German
	J. Janer et al (2006) [30]	Timing, F0, intensity, vibrato	Sample-based	Generic	Lyrics, MIDI notes, singing voice	Spanish
	T. Nakano et al (2009) [31]	Timing, F0, intensity	Sample-based	Popular Music database RWC[4]	Lyrics, singing voice	Japanese
	T. Nakano et al (2011) [32]	Timing, F0, intensity, timbre	Sample-based	Music Genre database in RWC	Lyrics, singing voice	Japanese
	T. Saitou et al (2007) [33]	Timing, F0, singer formant	Resynthesis of speech	Children's songs	Score, tempo, speech	Japanese
Rule-based	J. Sundberg (2006) [3]	Timing, consonant duration, vowel onset, timbre changes, formant tuning, overtone singing, articulation silence to note	Formant synthesis	Opera	Score, MIDI, or keyboard	Not specified
	M .Alonso (2005) [37]	Timing, micro-pauses, tempo and phrasing, F0, intensity, vibrato and tremolo, timbre quality	Sample-based	Angry, sad, happy	Score, lyrics, tempo, expressive intentions	Swedish, English
	J. Bonada (2008) [40]	Timbre (manual), phonetics, timing, F0, intensity, musical articulation, sustains, vibrato and tremolo (rate and depth)	Sample-based	Generic	Score, lyrics, tempo	Japanese, English, Spanish
Statistical modeling	K. Saino et al (2006) [25]	Timing, F0, timbre	HMM-based	Children's songs	Score and lyrics	Japanese
	K. Oura et al (2010) [42]	Timing, F0, vibrato and tremolo, timbre, source	HMM-based	Children's songs	MusicXML[5] score	Japanese, English
	K. Saino et al (2010) [22]	Baseline F0 (relative to note), vibrato rate and depth (not tremolo), intensity	Sample-based	Children's songs	Score (no lyrics to create models)	Japanese
Unit Selection	M. Umbert et al (2013) [43]	F0, vibrato, tremolo, intensity	Sample-based	Jazz standards	Score	Language independent

[4] https://staff.aist.go.jp/m.goto/RWC-MDB/

[5] http://www.musicxml.com/

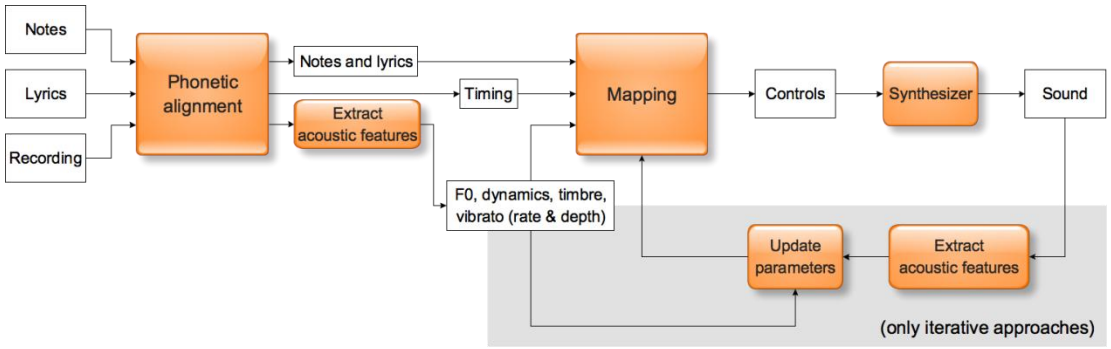


Fig. 5: General framework for performance-driven approaches.

Approaches using estimated controls achieve different levels of robustness depending on the singing voice synthesizers and voice databases. In the system presented in [29], a unit selection framework is used to create a singing voice synthesizer from a particular singer's recording in a nearly automatic procedure. In comparison to sample-based system, where the design criterion is to minimize the size of the voice database with only one possible unit sample (e.g. diphones), the criterion in unit selection is related to redundancy in order to allow the selection of consecutive units in the database, at the expense of having a larger database. The system automatically segments the recorded voice into phonemes by aligning it to the score and feeding the derived segmentation constraints to an HMM recognition system. Units are selected to minimize a cost function that scores the amount of time, frequency, and timbre transformations. Finally, units are concatenated. In this approach, the main effort is put on the synthesis engine. Although it uses a unit selection-based synthesizer, the expression controls for pitch, timing, dynamics, and timbre like the singer's formant are extracted from a reference singing performance of the target score. These parameters are directly used by the synthesizer to modify the selected units with a combination of sinusoidal modeling with PSOLA called SM-PSOLA. Editing is allowed by letting the user participate in the unit selection process, change some decisions, and modify the unit boundaries. Unfortunately, this approach only manipulates the singer's formant feature of timbre so that other significant timbre related features in opera singing style are not handled.

In [30], the followed steps are: extraction of acoustic features like energy, F0, and automatic

Table 5: Mapping from acoustic features to synthesizer controls.

	Mapped synthesis parameters				
Acoustic features	**Y. Meron (1999) [29]**	**J. Janer et al (2006) [30]**	**T. Nakano et al (2009) [31]**	**T. Nakano et al (2011) [32]**	**T. Saitou et al (2007) [33]**
F0	F0	Smoothed and continuous pitch	MIDI note number, pitch bend and sensitivity	MIDI note number, pitch bend and sensitivity	F0
Vibrato	Included in F0 implicitly	Vibratos from input or from DB singer	Included in F0 implicitly	Included in F0 implicitly	Included in F0 implicitly
Energy	Dynamics	Dynamics	Dynamics	Dynamics	Dynamics
Phonetic alignment	Phoneme timing	Onsets of vowels or voiced phonemes	Note onset and duration	Note onset and duration	Phoneme timing
Timbre	Singer's formant cluster amplitude	Not used	Not used	Mixing different voice quality DBs	Singer's formant cluster amplitude and AM of the synthesized signal

detection of vibrato sections, mapping into synthesis parameters, and phonetic alignment. The mapped controls and the input score are used to build an internal score that matches the target timing, pitch, and dynamics, and minimizes the transformation cost of samples from a database. However, this approach is limited since timbre is not handled and also because the expression features of the synthesized performance are not compared to the input values. Since this approach lacks a direct mapping of acoustic features to control parameters, these differences are likely to happen. On the other hand, the possibility of using a singer DB to produce vibratos other than the extracted ones from the reference recording provides a new degree of freedom to the user.

Toward a more robust methodology to estimate the parameters, in [31] the authors study an iterative approach that takes the target singing performance and lyrics as. The musical score or note sequence is automatically generated from the input. The first iteration provides an initialization of the system similar to the previous approach [30]. At this point these controls can be manually edited by applying pitch transposition, correction, vibrato modifications, and pitch and intensity smoothing. The iterative process continues by analyzing the synthesized waveform and adjusting the control parameters so that in the next iteration the results are closer to the expected performance. In [32], the authors extend this approach by including timbre. Using different voice quality databases from the same singer, the corresponding versions of the target

song are synthesized as in the previous approach. The system extracts the spectral envelopes of each one to build a 3-dimensional voice timbre space. Next, a temporal trajectory in this space is estimated from the reference target performance in order to represent its spectral timbre changes. Finally, singing voice synthesis output is generated using the estimated trajectory to imitate the target timbre change. Although expression control is more robust than the previous approach thanks to iteratively updating the parameters and by allowing a certain degree of timbre control, these approaches also have some limitations. First, it cannot be assured that the iterative process will converge to the optimal set of parameter values. Secondly, the timbre control is limited to the variability within the set of available voice quality databases.

In [33], naturally-spoken readings of the target lyrics are transformed into singing voice by matching the target song properties described in the musical score. Other input data are the phonetic segmentation and the synchronization of phonemes and notes. The approach first extracts acoustic features like F0, spectral envelope, and the aperiodicity index from the input speech. Then, a continuous F0 contour is generated from discrete notes, phoneme durations are lengthened, and the singer's formant cluster is generated. The fundamental frequency contour takes into account four types of fluctuations, namely, overshoot (F0 exceeds the target note after a note change), vibrato, preparation (similar to overshoot before the note change), and fine fluctuations. The first three types of F0 fluctuations are modeled by a single second-order transfer function that depends mainly on a damping coefficient, a gain factor and a natural frequency. A rule-based approach is followed for controlling phoneme durations by splitting consonant-to-vowel transitions into three parts. First, the transition duration is not modified for singing. Then, the consonant part is transformed based on a comparative study of speech and singing voices. Finally, the vowel section is modified so that the duration of the three parts matches the note duration. Finally, with respect to timbre, the singer's formant cluster is handled by an emphasis function in the spectral domain centered at 3 kHz. Amplitude modulation is also applied to the synthesized singing voice according to the generated vibratos parameters. Although we have

classified this approach into the performance-driven section since the core data is found in the input speech recording, some aspects are modeled like the transfer function for F0, rules for phonetic duration, and a filter for the singer's formant cluster. Similarly to [29], in this approach timbre control is limited to the singer formant, so that the system cannot change other timbre features. However, if the reference speech recording contains voice quality variations that fit the target song, this can add some naturalness to the synthesized singing performance.

Performance-driven approaches achieve a highly expressive control since performances implicitly contain knowledge naturally applied by the singer. These approaches become especially convenient for creating parallel database recordings which are used in voice conversion approaches [8]. On the other hand, the phonetic segmentation may cause timing errors if not manually corrected. The non-iterative approach lacks robustness because the differences between input controls and the extracted ones from the synthesized sound are not corrected. In [32] timbre control is limited by the number of available voice qualities. We note that a human voice input for natural singing control is required for these approaches, which can be considered as a limitation since it may not be available in most cases. When such a reference is not given, other approaches are necessary to derive singing control parameters from the input musical score.

D. Rule-based approaches

Rules can be derived from work with synthesizing and analyzing sung performances. Applying an analysis-by-synthesis method an ambitious rule-based system for Western music was developed at KTH in the 1970s and improved over the last three decades [3]. By synthesizing sung performances, this method aims at identifying acoustic features that are perceptually important either individually or jointly [15]. The process of formulating a rule is iterative. First a tentative rule is formulated and implemented and the resulting synthesis is assessed. If its effect on the performance needs to be changed or improved, the rule is modified and the effect of the resulting performance is again assessed. On the basis of parameters such as phrasing, timing,

Table 6: Singing voice related KTH rules' dependencies.

Acoustic feature	Dependencies
Consonant duration	Previous vowel length
Vowel onset	Synchronized with timing
Formant frequencies	Voice classification
Formant frequencies	Pitch, if otherwise F0 would exceed the first formant
Spectrum slope	Decrease with increasing intensity
Vibrato	Increase depth with increasing intensity
Pitch in coloratura passages	Each note represented as a vibrato cycle
Pitch phrase attack (and release)	At pitch start (end) from (at) 11 semitones below target F0

metrics, note articulation, and intonation, the rules modify pitch, dynamics, and timing. Rules can be combined to model emotional expressions as well as different musical styles. Table 6 lists some of the acoustic features and their dependencies.

The rules reflect both physical and musical phenomena. Some rules are compulsory and others optional. The *Consonant duration* rule, which lengthens consonants following short vowels, applies also to speech in some languages. The *Vowel onset* rule corresponds to the general principle that the vowel onset is synchronized with the onset of the accompaniment, even though lag and lead of onset are often used for expressive purposes [34]. The *Spectrum slope* rule is compulsory, as it reflects the fact that vocal loudness is controlled by subglottal pressure and an increase of this pressure leads to a less steeply sloping spectrum envelope. The rule *Pitch in coloratura passages* implies that the fundamental frequency makes a rising-falling gesture around the target frequency in legato sequences of short notes [35]. The *Pitch phrase attack*, in the lab jargon referred as the "Bull's roaring onset", is an ornament used in excited moods, and would be completely out of place in a tender context. Interestingly, results close to the KTH rules have been confirmed by machine learning approaches [36].

A selection of the KTH rules [15] has been applied to the Vocaloid synthesizer [37]. Features are considered at note level (start and end times), intra and inter note (within and between note changes) and to timbre variations (not related to KTH rules). The system implementation is detailed in [38], along with the acoustic cues which are relevant for conveying basic emotions

such as anger, fear, happiness, sadness, and love-tenderness [12]. The rules are combined in expressive palettes indicating to what degree rules need to be applied to convey a target emotion. The relationship between application level, rules, and acoustic features is shown in Table 7. As an example of the complexity of the rules, the punctuation rule at note level inserts a 20 milliseconds micro-pause if a note is three tones lower than the next one and its duration is 20% larger. Given that this work uses a sample-based synthesizer, voice quality modifications are applied to the retrieved samples. In this case, the timbre variations are limited to rules affecting brightness, roughness, and breathiness, and therefore do not cover the expressive possibilities of a real singer.

Apart from the KTH rules, in corpus-derived rule-based systems heuristic rules are obtained to control singing expression by observing recorded performances. In [6], expression controls are generated from high-level performance scores where the user specifies note articulation, pitch, intensity, and vibrato data which is used to retrieve templates from recorded samples. This work, used in the Vocaloid synthesizer [39], models the singer's performance with heuristic rules [40]. The parametric model is based on anchor points for pitch and intensity, which are manually derived from the observation of a small set of recordings. At synthesis, the control contours are obtained by interpolating the anchor points generated by the model. The number of points used for each note depends on its absolute duration. The phonetics relationship with timing is handled by synchronizing the vowel onset with the note onset. Moreover, manual editing is permitted for the degree of articulation application as well as its duration, pitch and dynamics contours, phonetic transcription, timing, vibrato and tremolo depth and rate, and timbre characteristics.

The advantage of these approaches is that they are relatively straight-forward and completely deterministic. Random variations can be easily introduced so that the generated contours are different for each new synthesis of the same score, resulting in distinct interpretations. The main drawbacks are that either the models are based on few observations that do not fully represent a given style, or they are more elaborate but become unwieldy due to the complexity of the rules.

Table 7: Selection of rules for singing voice: level of application and affected acoustic features.

Level	Rules	Affected acoustic features
Note	Duration contrast	Decrease duration and intensity of short notes placed next to long notes
	Punctuation	Insert micro-pauses in certain pitch interval and duration combinations
	Tempo	Constant value for the note sequence (measured in bpm)
	Intensity	Smooth/strong energy levels, high pitch notes intensity increases 3 dB/octave
	Transitions	Legato, staccato (pause is set to more than 30% of inter-onset interval)
	Phrasing arch	Increase/decrease tempo at phrase beginning/end, same for energy
	Final ritardando	Decrease tempo at the end of a piece
Intra/Inter note	Attack	Pitch shape from starting pitch until target note, energy increases smoothly
	Note articulation	Pitch shape from the starting to the ending note, smooth energy
	Release	Energy decreases smoothly to 0, duration is manually edited
	Vibrato and tremolo	Manual control of position, depth, and rate (cosine function, random fluctuations)
Timbre	Brightness	Increase high frequencies depending on energy
	Roughness	Spectral irregularities
	Breathiness	Manual control of noise level (not included in emotion palettes)

E. Statistical modeling approaches

Several approaches have been used to statistically model and characterize expression control parameters using Hidden Markov Models (HMMs). They have a common precedent in speech synthesis [41], where the parameters like spectrum, F0 and state duration are jointly modeled. Compared to unit selection, HMM-based approaches tend to produce lower speech quality, but they need a smaller dataset to train the system without needing to cover all combinations of contextual factors. Modeling singing voice with HMMs amounts to using similar contextual data as for speech synthesis, adapted to singing voice specificities. Moreover, new voice characteristics can be easily generated by changing the HMM parameters.

These systems operate in two phases: training and synthesis. In the training part, acoustic features are first extracted from the training recordings like F0, intensity, vibrato parameters, and mel-cepstrum coefficients. Contextual labels, that is to say, the relationships of each note, phoneme, phrase with the preceding and succeeding values, are derived from the corresponding score and lyrics. Contextual labels vary in their scope at different levels, such as phoneme, note, or phrase, according to the approach, as summarized in Table 8. This contextual data is used to

Table 8: Contextual factors HMM-based systems (P/C/N stands for: Previous, Current, and Next).

HMM-based approaches	Levels	Contextual factors
K. Saino et al (2006) [25]	Phoneme	P/C/N phonemes
	Note	P/C/N note F0, durations, and positions within the measure
K. Oura et al (2010) [42]	Phoneme	Five phonemes (central and two preceding and succeeding)
	Mora	Number of phonemes in the P/C/N mora
		Position of the P/C/N mora in the note
	Note	Musical tone, key, tempo, length, and dynamics of the P/C/N note
		Position of the current note in the current measure and phrase
		Ties and slurred articulation flag
		Distance between current note and next/previous accent and staccato
		Position of the current note in the current crescendo or decrescendo
	Phrase	Number of phonemes and moras in the P/C/N phrase
	Song	Number of phonemes, moras, and phrases in the song
K. Saino et al (2010) [22]	Note region	Manually segmented behaviour types (beginning, sustained, ending)
	Note	MIDI note number and duration (in 50 ms units)
		Detuning: model F0 by the relative difference to the nominal note

build the HMMs that relate how these acoustic features behave according to the clustered contexts. The phoneme timing is also modeled in some approaches. These generic steps for the training part in HMM-based synthesis are summarized in Fig. 6. The figure shows several blocks found in the literature, which might not be present simultaneously in each approach. We refer to [41] for the detailed computations that HMM training involves.

In the synthesis part, given a target score, contextual labels are derived as in the training part from the note sequence and lyrics. Models can be used in two ways. All necessary parameters forsinging voice synthesis can be generated from them, therefore state durations, F0, vibrato and mel-cepstrum observations are generated to synthesize the singing voice. On the other hand, if another synthesizer is used, only control parameters such as F0, vibrato depth and rate, and dynamics need to be generated which are then used as input of the synthesizer.

As introduced in Section III.A, statistical methods jointly model the largest set of expression features among the reviewed approaches. This gives them a better generalization ability. As long as singing recordings for training involve different voice qualities, singing styles or emotions, and the target language phonemes, these will be reproducible at synthesis given the appropriate

context labeling. Model interpolation allows new models to be created as a combination of existing ones. New voice qualities can be created by modifying the timbre parameters. However, this flexibility is possible at the expense of having enough training recordings to cover the combinations of the target singing styles and voice qualities. In the simplest case, a training database of a set of songs representing a single singer and style in a particular language would be enough to synthesize it. As a drawback, training HMMs with large databases tends to produce smoother time series than the original training data, which may be perceived as non-natural.

In [25], a corpus-based singing voice synthesis system based on HMMs is presented. The contexts are related to phonemes, note F0 values, and note durations and positions, as we show in Table 8 (dynamics are not included). Also, synchronization between notes and phonemes needs to be handled adequately, mainly because phoneme timing does not strictly follow the score timing; and phonemes might be advanced with respect to the nominal note onsets (negative time-lag).

In this approach, the training part generates three models. One for the spectrum where MFCCs are estimated with STRAIGHT and excitation (F0) parts, extracted from the training database, another for the duration of context-dependent states, and a third one to model the time-lag. The latter ones model note timing and phoneme durations of real performances, which are different to what can be inferred from the musical score and its tempo. Time-lags are obtained by forced alignment of the training data with context-dependent HMMs. Then, the computed time-lags are related to their contextual factors and clustered by a decision-tree.

The singing voice is synthesized in five steps. First, the input score (note sequence and lyrics) is analyzed to determine note duration and contextual factors. Then, a context-dependent label sequence of contextual factors as shown in Table 8 is generated. Then, the song HMM is generated and its state durations are jointly determined with the note time-lags. Next, spectral and F0 parameters are generated, which are used to synthesize the singing voice. The authors claim that the synthesis performance achieves a natural singing voice which simulates expression elements of the target singer such as voice quality and the singing style (F0 and time-lag).

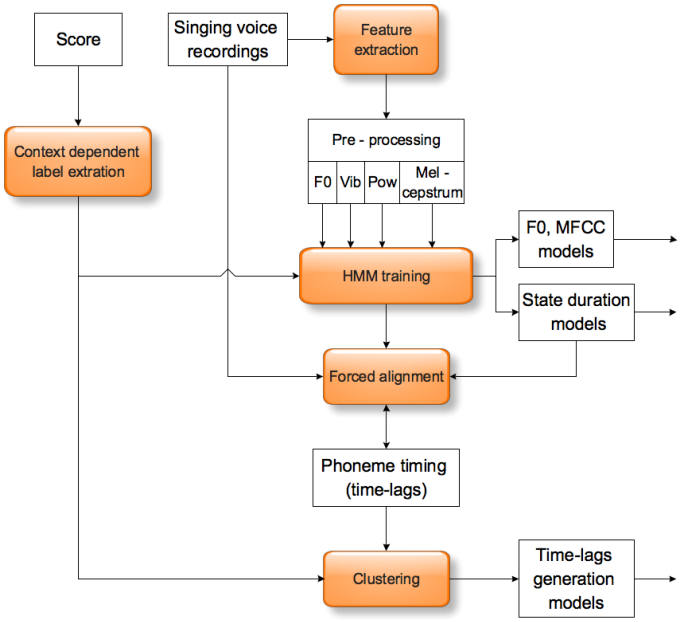


Fig. 6: Generic blocks for the training part of HMM-based approaches.

In this work, the training database consists of 72 minutes of a male voice singing 60 Japanese children's songs in a single voice quality. These are the characteristics that the system can reproduce in a target song. The main limitation of this approach is that contextual factors scope is designed only to cover phoneme and note descriptors. Longer scopes than just the previous and next note are necessary to model higher level expressive features such as phrasing. Although we could not get samples from this work, an evolved system is presented next.

The system presented in [25] has been improved, and is publicly available as Sinsy, an online singing voice synthesizer [42]. The new characteristics of the system include reading input files in MusicXML format[6] with F0, lyrics, tempo, key, beat, and dynamics, also extended contextual factors used in the training part, vibrato rate and depth modeling, and a reduction of the computational cost. Vibrato is jointly modeled with the spectrum and F0 by including depth and rate in the observation vector in the training step.

The new set of contexts, automatically extracted from the musical score and lyrics, used by the Sinsy approach are also shown in Table 8. These factors describe the context such as previous,

[6] http://www.musicxml.com/

current, and next data at different hierarchical levels, namely, phoneme, mora (the sound unit containing one or two phonemes in Japanese), note, phrase, and the entire song. Some of them are strictly related to musical expression aspects, such as musical tone, key, tempo, length and dynamics of notes, articulation flags, or distance to accents and staccatos.

Similarly to the previous work, in this case the training database consists of 70 minutes of a female voice singing 70 Japanese children's songs in a single voice quality. However, it is able to reproduce more realistic expression control since vibrato parameters are also extracted and modeled. Notes are described with a much richer set of factors than the previous work. Another major improvement is the scope of the contextual factors shown in Table 8, which spans from the phoneme level up to the whole song and therefore being able to model phrasing.

In [22], a statistical method is able to model singing styles. This approach focuses on baseline F0, vibrato features like its extent, rate, and evolution over time, not tremolo, and dynamics. These parameters control the Vocaloid synthesizer, and so timbre is not controlled by the singing style modeling system, but is dependent on the database.

A preprocessing step is introduced after extracting the acoustic features like F0 and dynamics in order to get rid of the micro-prosody effects on such parameters, by interpolating F0 in unvoiced sections and flattening F0 valleys of certain consonants. The main assumption here is that expression is not affected by phonetics, which is reflected in erasing such dependencies in the initial preprocessing step, and also in training note HMMs instead of phoneme HMMs. Also, manual checking is done to avoid errors in F0 estimation and MIDI events like note on and note off estimated from the phonetic segmentation alignment. A novel approach estimates vibrato shape and rate, which at synthesis is added to the generated baseline melody parameter. The shape is represented with the low frequency bins of the Fourier Transform of single vibrato cycles. In this approach, context-dependent HMMs model the expression parameters which are summarized in Table 8. Feature vectors contain melody, vibrato shape and rate, and dynamics components.

This last HMM-based work focuses on several control features except timbre, which is handled by the Vocaloid synthesizer. This makes the training database much smaller in size. It consists of 5 minutes of 5 Japanese children's songs, since there is no need to cover a set of phonemes. Contextual factors are rich at a note level, since the notes are divided into 3 parts (begin, sustain, and end), and the detuning is also modeled relatively to the nominal note. On the other hand, this system lacks of the modeling of wider temporal aspects such as phrasing.

F. Unit selection approaches

The main idea of unit selection [29] is to use a database of singing recordings segmented into units which consist of one or more phonemes or other units like diphones or half phones. For a target score, a sequence of phonemes with specific features like pitch or duration is retrieved from the database. These are generally transformed to match the exact required characteristics.

An important step in this kind of approach is the definition of the target and concatenation cost functions as the criteria on which unit selection is built. The former is a distance measure of the unit transformation in terms of a certain acoustic feature like pitch, duration, etc. Concatenation costs measure the perceptual consequences of joining non-consecutive units. These cost functions' contributions are weighted and summed to get the overall cost of the unit sequence. The goal is then to select the sequence with the lowest cost.

Unit selection approaches present the disadvantages of requiring a large database, which needs to be labeled, and that subcost weights need to be determined. On the other hand, the voice quality and naturalness are high due to the implicit rules applied by the singer within the units.

A method to model pitch, vibrato features, and dynamics based on selecting units from a database of performance contours has recently been proposed [43]. We illustrate it in Fig. 7 for the F0 contour showing two selected source units for a target note sequence where units are aligned at the transition between the 2^{nd} and 3^{rd} target notes. The target note sequence is used as input to generate the pitch and dynamics contours. A reference database is used, containing

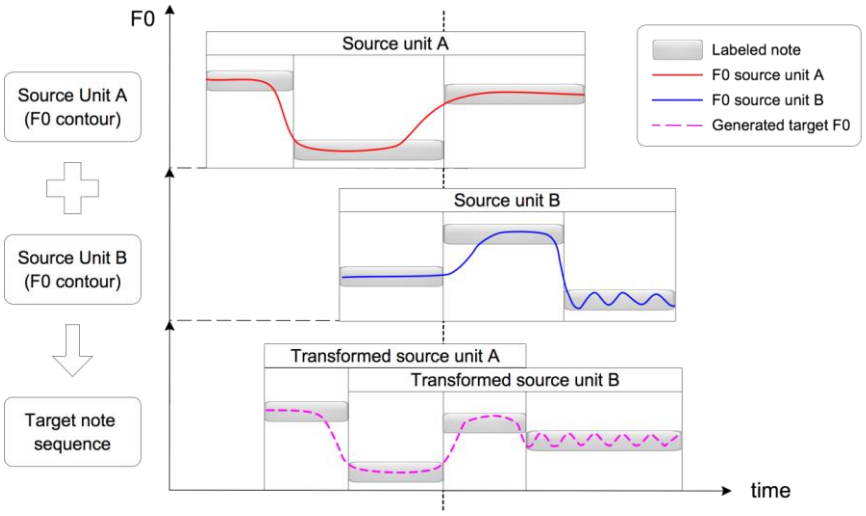


Fig. 7: Performance feature (F0) generated by unit selection.

extracted pitch, vibrato features, and dynamics from expressive recordings of a single singer and style. Besides these features, the database is labeled with the note pitches, durations, strength, and the start and end times of note transitions. This approach splits the task of generating the target song expression contours into first finding similar and shorter note combinations (source units A and B in Fig. 7), and then transforming and concatenating the corresponding pitch and dynamics contours in order to match the target score (dashed line in Fig. 7). These shorter contexts are the so-called units, defined by three consecutive notes or silences, so that consecutive units overlap by two notes. The contour of dynamics is generated similarly from the selected units.

With regard to unit selection, the cost criterion consists of the combination of several sub-cost functions, as summarized in Table 9. In this case, there are four functions and unit selection is implemented with the Viterbi algorithm. The overall cost function considers the amount of transformation in terms of note durations (note duration cost) and pitch interval (pitch interval cost), in order to preserve as much as possible the contours as originally recorded. It also measures how appropriate it is to concatenate two units (concatenation cost), as a way of penalizing the concatenation of units from different contexts. Finally, the overall cost function also favors the selection of long sequences of consecutive notes (continuity cost), although the final number of consecutive selected units depends on the resulting cost value. This last

Table 9: Unit selection cost functions.

Cost	Description	Computation
Note duration	Compare source and target unit note durations	Octave ratio (source/target unit notes)
Pitch interval	Compare source and target unit note intervals	Octave ratio (source/target unit intervals)
Concatenation	Favour compatible units from the DB	0 if consecutive units
Continuity	Favour selection of consecutive units	Penalize selection of non-consecutive units

characteristic is relevant in order to be able to reflect the recorded phrasing at synthesis.

Once a sequence is retrieved, each unit is time scaled and pitch shifted. The time scaling is not linear, instead most of the transformation is applied in the sustain part and keeping the transition (attacks and releases) durations as close to the original as possible. Vibrato is handled with a parametric model, which allows the original rate and depth contour shapes to be kept.

The transformed unit contours are overlapped and added after applying a crossfading mask, which mainly keeps the shape of the attack to the unit central note. This is done separately for the intensity, baseline pitch and vibrato rate, and vibrato depth contours. The generated baseline pitch is then tuned to the target note pitches in order to avoid strong deviations. Then vibrato rate and depth contours are used to compute the vibrato oscillations which are added to the baseline pitch. Concerning the expression database, it contains several combinations of note durations, pitch intervals, and note strength. Such a database can be created systematically [44] in order to cover a relevant portion of possible units. Notes are automatically detected and then manually checked. Vibrato sections are manually segmented and depth and rate contours are estimated. An important characteristic of such database is that it does not contain sung text, only sung vowels to avoid micro-prosody effects when extracting pitch and dynamics.

This approach controls several expression features except timbre aspects of the singing voice. In our opinion, a positive characteristic is that it can generate expression features without suffering from smoothing as is the case in HMMs. The selected units contain the implicit rules applied by the singer in order to perform a vibrato, an attack, or a release. Besides, the labeling and cost functions for unit selection are designed in a way that favors the selection of long

sequences of consecutive notes in the database to help the implicit reproduction of high expression features like phrasing. Similarly to the KTH rules, this approach is independent of phonetics since this is handled separately by the controlled synthesizer, which makes it convenient for any language. The lack of an explicit timbre control could be addressed in the future by adding control features like the degree of breathiness or brightness.

In the previous subsections we have classified, compared, described, and analyzed each type of approach. In the next subsection we provide an insight on when to use each approach.

G. When to use each approach?

The answer to this question has several considerations: from the limitations of each approach, to whether singing voice recordings are available or not since these are needed in model training or unit selection, the reason for synthesizing a song which could be for database creation or rule testing, or flexibility requirements like model interpolation. In this section we provide a brief guideline on the suitability of each type of approach.

Performance-driven approaches are suitable to be applied, by definition, when the target performance is available, since the expression of the singer is implicit in the reference audio and it can be used to control the synthesizer. Another example of applicability is the creation of parallel databases for different purposes like voice conversion [8]. An application example for the case of speech to singing synthesis is the generation of singing performances for untrained singers, whose timbre is taken from the speech recording and the expression for pitch and dynamics can be obtained from a professional singer.

Rule-based approaches are suitable to be applied to verify the defined rules and also to see how these are combined, for example to convey a certain emotion. If no recordings are available, rules can still be defined with the help of an expert, so that these approaches are not fully dependent on singing voice databases.

Statistical modeling approaches are also flexible, given that it is possible to interpolate models

and to create new voice characteristics. They have the advantage that in some cases these are part of complete singing voice synthesis systems, that is to say, the ones that have the score as input and that generate both the expression parameters and output voice.

Similarly to rule-based and statistical modeling approaches, unit selection approaches do not need the target performance, although they can benefit from it. On the other hand, unit selection approaches share a common characteristic with performance-driven approaches. The implicit knowledge of the singer is contained in the recordings, although in unit selection it is extracted from shorter audio segments. Unlike statistical models, no training step is needed, so that the expression databases can be improved just by adding new labeled singing voice recordings.

In the following section we review the evaluation strategies of the expression control approaches, identify some deficiencies, and finally propose a possible solution.

IV. Evaluation

In Section I, we introduced that a score can be interpreted in several acceptable ways, which makes expression a subjective aspect to rate. However, "*procedures for systematic and rigorous evaluation do not seem to exist today*" [1] (p. 105), especially if there is no ground-truth to compare with. In this section, we first summarize typical evaluation strategies. Then, we propose the initial ideas to build a framework that solves some detected issues, and finally we discuss the need for automatic measures to rate expression.

A. Current evaluation strategies

Expression control can be evaluated from subjective or objective perspectives. The former typically consists of listening tests where participants perceptually evaluate some psychoacoustic characteristic like voice quality, vibrato, and overall expressiveness of the generated audio files. A common scale is the mean opinion score (MOS), with a range from 1 (bad) to 5 (good). In pair-wise comparisons, using two audio files obtained with different system configurations, preference tests rate which option achieves a better performance. Objective evaluations help to compare how

Table 10: Conducted subjective and objective evaluations per approach.

		Tests		
Type	**Approach**	**Method**	**Description**	**Subjects**
Performance-driven	Y. Meron (1999) [29]	Subjective	Rate voice quality with pitch modification of 10 pairs of sentences (SM-PSOLA vs TD-PSOLA)	10 subjects
	J. Janer et al (2006) [30]	Subjective	Informal listening test	Not specified
	T. Nakano et al (2009) [31]	Objective	Two tests: lyrics alignment and mean error value of each iteration for F0 and intensity compared to target	No subjects
	T. Nakano et al (2011) [32]	Objective	Two tests: 3D voice timbre representation and Euclidean distance between real and measured timbre	No subjects
	T. Saitou (2007) [33]	Subjective	Paired comparisons of different configurations to rate naturalness of synthesis in a 7 step scale (-3 to 3)	10 students with normal hearing ability
Rule-based	J. Sundberg (2006) [3]	Subjective	Listening tests of particular acoustic features	15 singers or singing teachers
	M. Alonso (2005) [37]	None	None	None
	J. Bonada (2008) [40]	Subjective	Listening tests ratings (1-5)	50 subjects with different levels of musical training
Statistical modelling	K. Saino et al (2006) [25]	Subjective	Listening test (1-5 ratings) of 15 musical phrases. Two tests: with and without time-lag model	14 subjects
	K. Oura et al (2010) [42]	Subjective	Not detailed (based on Saino 2006)	Not specified
	K. Saino et al (2010) [22]	Subjective	Rate style and naturalness listening tests ratings (1-5) of 10 random phrases per subject	10 subjects
Unit selection	M. Umbert et al (2013) [43]	Subjective	Rate expression, naturalness, and singer skills listening tests ratings (1-5)	17 subjects with different levels of musical training

well the generated expression controls match a reference real performance by computing an error. Within the reviewed works, subjective tests outnumber the objective evaluations. In Table 10 the evaluations are summarized. For each approach, several details are provided like a description of the evaluation (style, voice quality, naturalness, expression, and singer skills), the different rated tests, and information on the subjects if available. Objective tests are done only for performance-driven approaches, that is to say, when a ground-truth is available. In the other approaches, no reference is directly used for comparison, so that only subjective tests are carried out. However, in the absence of a reference of the same target song, the generated performances could be compared to the recording of another song, as is done in the case of speech synthesis.

In our opinion, the described evaluation strategies are devised for evaluating a specific system, and therefore focus on a concrete set of characteristics particularly relevant for that system. For instance, the evaluations summarized in Table 10 do not include comparisons to other

approaches. This is due to the substantial differences between systems, which make the evaluation and comparison between them a complex task. These differences can be noted in the audio excerpts of the accompanying website to this article, which have been introduced in Section III.B. At this stage, it is difficult to decide which method more efficiently evokes a certain emotion or style, performs better vibratos, changes the voice quality in a better way, or has a better timing control. There are limitations in achieving such a comprehensive evaluation and comparing the synthesized material. In the next section we propose a possible solution.

B. Towards a common evaluation framework

The evaluation methodology could be improved by building the systems under similar conditions to reduce the differences among performances and by sharing the evaluation criteria. Building a common framework would help to easily evaluate and compare the singing synthesis systems.

The main blocks of the reviewed works are summarized in Fig. 8. For a given target song, the expression parameters are generated to control the synthesis system. In order to share as many commonalities as possible amongst systems, these could be built under *similar conditions* and tested by a shared *evaluation criterion*. Then, the comparison would benefit from focusing on the technological differences and not on other aspects like the target song and singer databases.

Concerning the *conditions*, several aspects could be shared amongst approaches. Currently, there are differences in the target songs synthesized by each approach, the set of controlled expression features, and the singer recordings (e.g. singer gender, style, or emotion) used to derive rules, to train models, to build expression databases, and to build the singer voice models.

A publicly available dataset of songs, with both scores (e.g. in MusicXML format) and reference recordings, could be helpful if used as target songs in order to evaluate how expression is controlled by each approach. In addition, deriving the expression controls and building the voice models from a common set of recordings would have a great impact on developing this evaluation framework. If all approaches shared such a database, it would be possible to compare

how each one captures expression and generates the control parameters, since the starting point would be the same for all them. Besides, both sample-based and HMM-based synthesis systems would derive from the same voice. Thus, it would be possible to test a single expression control method with several singing voice synthesis technologies. The main problem we envisage is that some approaches are initially conceived for a particular synthesis system. This might not be a major problem for the pitch contour control, but it would be more difficult to apply the voice timbre modeling of HMM-based systems to sample-based systems.

The subjective evaluation process is worthy of particular note. Listening tests are a time consuming task and several aspects need to be considered in their design. The different backgrounds related to singing voice synthesis, speech synthesis, technical skills, and the wide range of musical skills of the selected participants can be taken into consideration by grouping the results according to such expertise, and clear instructions have to be provided on what to rate like to focus on specific acoustic features of the singing voice, and how to rate using pair-wise comparisons or MOS. Moreover, uncontrolled biases in the rating of stimuli due to the order in which these are listened can be avoided by presenting them using pseudo-random methods like Latin-squares, and the session duration has to be short enough to not decrease the participant's level of attention. However, often the reviewed evaluations have been designed differently and are not directly comparable. In the next section, we introduce a proposal to overcome this issue.

C. Perceptually-motivated objective measures

The constraints in Section IV.B make unaffordable to extensively evaluate different configurations of systems by listening to many synthesized performances. This could be solved if objective measures that correlate with perception were established. Such perceptually-motivated objective measures could be computed by learning the relationship between MOS and extracted features at a local or global scope. The measure should be ideally independent from the style and the singer, and it should provide ratings for particular features like timing, vibratos, tuning, voice

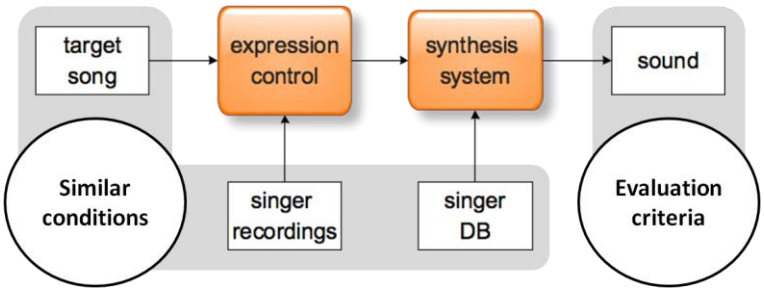


Fig. 8: Proposed common evaluation framework.

quality, or the overall performance expression. These measures, besides helping to improve the systems' performance, would represent a standard for evaluation and allow for scalability.

The development of perceptually-motivated objective measures could benefit from approaches in the speech and audio processing fields. Psychoacoustic and cognitive models have been used to build objective metrics for assessing audio quality and speech intelligibility [45] and its effectiveness has been measured by its correlation to MOS ratings. Interestingly, method specific measures have been computed in unit selection cost functions for speech synthesis [46]. Other approaches for speech quality prediction are based on a log-likelihood measure as a distance between a synthesized utterance and an HMM model built from features based on MFCCs and F0 of natural recordings [47]. This gender-dependent measure is correlated to subjective ratings like naturalness. For male data, it can be improved by linearly combining it with parameters typically used in narrow-band telephony applications, like noise or robotization effects. For female data, it can be improved by linearly combining it with parameters related to signal like duration, formants, or pitch. The research on automatic evaluation of expressive performances is considered an area to exploit, although it is still not mature enough [48], for example, it could be applied to develop better models and training tools for both systems and students.

Similarly to the speech and instrumental music performance communities, the progress in the singing voice community could be incentivized through evaluation campaigns. These types of evaluations help to identify the aspects that need to be improved and can be used to validate perceptually-motivated objective measures. Examples of past evaluation campaigns are the

Synthesis Singing Challenge[7] and the Performance Rendering Contest[8] (Rencon) [48]. In the first competition, one of the target songs was compulsory and the same for each team. Performances were rated by 60 participants with a five-point scale involving quality of the voice source, quality of the articulation, expressive quality, and the overall judgment. The organizers concluded "*the audience had a difficult task, since not all systems produced both a baritone and a soprano version, while the quality of the voices used could be quite different (weaker results for the female voice)*"[7]. The Rencon's methodology is also interesting. Expressive performances are generated from the same Disklavier grand piano, so that the differences among approaches are only due to the performance and subjectively evaluated by an audience and experts. In 2004, voice synthesizers were also invited. Favorable reviews were received but not included in the ranking.

In this section we have seen challenges related to the evaluation process like the common framework for the evaluation and perceptually-motivated objective measures. In the next section, we identify and discuss other challenges not strictly related to the evaluation.

V. Challenges

While expression control has advanced in recent years, there are many open challenges. In this section, we discuss some specific challenges and consider the advantages of hybrid approaches. Next, we discuss important challenges in approaching a more human-like naturalness in the synthesis. Then, requirements for intuitive and flexible singing voice synthesizers' interfaces are analyzed, as well as the importance of associating a synthetic voice with a character.

A. Towards hybrid approaches

Several challenges have been identified in the described approaches. Only one of the performance-driven approaches deals with timbre, and it depends on the available voice quality databases. This approach would benefit from techniques for the analysis of the target voice quality, its evolution over time, and techniques for voice quality transformations so to be able to

[7] http://www.interspeech2007.org/Technical/synthesis_of_singing_challenge.php
[8] http://renconmusic.org/

synthesize any type of voice quality. The same analysis and transformation techniques would be useful for the unit selection approaches. Rule-based approaches would benefit from machine learning techniques that learn rules from singing voice recordings in order to characterize a particular singer and to explore how these are combined. Statistical modeling approaches are currently not dealing with comprehensive databases that cover a broad range of styles, emotions, and voice qualities. If we could take databases that efficiently cover different characteristics of a singer in such a way, it would lead to interesting results like model interpolation.

We consider the combination of existing approaches to have great potential. Rule-based techniques could be used as a pre-preprocessing step to modify the nominal target score so that it contains variations such as ornamentations and timing changes related to the target style or emotion. The resulting score could be used as the target score for statistical and unit selection approaches where the expression parameters would be generated.

B. More human-like singing synthesis

One of the ultimate goals of singing synthesis technologies is to synthesize human-like singing voices that cannot be distinguished from human singing voices. Although the naturalness of synthesized singing voices has been increasing, perfect human-like naturalness has not yet been achieved. Singing synthesis technologies will require more dynamic, complex, and expressive changes in the voice pitch, loudness, and timbre. For example, voice quality modifications could be related to emotions, style, or lyrics.

Moreover, automatic context-dependent control of those changes will also be another important challenge. The current technologies synthesize words in the lyrics without knowing their meanings. In the future, the meanings of the lyrics could be reflected in singing expressions as human singers do. Human-like singing synthesis and realistic expression control may be a highly challenging goal, given how complex this has been proven for speech.

When human-like naturalness increases, the *"Uncanny Valley"* hypothesis [49] states that

some people may feel a sense of creepiness. Although the Uncanny Valley is usually associated with robots and computer graphics, it is applicable even to singing voices. In fact, when a demonstration video by VocaListener [31] first appeared in 2008, the Uncanny Valley was often mentioned by listeners to evaluate its synthesized voices. An exhibition of a singer robot driven by VocaWatcher [50] in 2010 also elicited more reactions related to the Uncanny Valley. However, we believe that such discussion of the Uncanny Valley should not discourage further research. What this discussion means is that the current technologies are in a transitional stage towards future technologies that will go beyond the Uncanny Valley [50], and that it is important for researchers to keep working towards such future technologies.

Note, however, that human-like naturalness is not always demanded. As sound synthesis technologies are often used to provide artificial sounds that cannot be performed by natural instruments, synthesized singing voices that cannot be performed by human singers are also important and should be pursued in parallel, sometimes even for aesthetic reasons. Some possible examples are extremely fast singing, or singing with pitch or timbre quantization.

C. More flexible interfaces for singing synthesis

User interfaces for singing synthesis systems will play a more important role in the future. As various score-driven and performance-driven interfaces are indispensable for musicians in using general sound synthesizers, singing synthesis interfaces have also had various options such as score-driven interfaces based on the piano-roll or score editor, and performance-driven interfaces in which a user can just sing along with a song and a synthesis system then imitates him or her (as mentioned in III.C.). More intuitive interfaces that do not require time-consuming manual adjustment will be an important goal for ultimate singing interfaces. So far, direct manipulator-style interfaces such as the above score-driven or performance-drive interfaces are used for singing synthesis systems, but indirect producer-style interfaces, such as those that enable users to verbally communicate with and ask a virtual singer to sing in different ways, will also be

attractive to help users focus on how to express the user's message or intention through a song, though such advanced interfaces have yet to be developed. More flexible expression control of singing synthesis in real-time is also another challenge.

D. Multimodal aspects of singing synthesis

Attractive singing synthesis itself must be a necessary condition for its popularity, but not a sufficient condition. The most famous virtual singer, *Hatsune Miku*, has shown that having a character can be essential to make singing synthesis technologies popular. Hatsune Miku is the name of the most popular singing synthesis software package in the world. She is based on Vocaloid and has a cute synthesized voice in Japanese and English with an illustration of a cartoon girl. After Hatsune Miku originally appeared in 2007, many people started listening to a synthesized singing voice as the *main vocal* of music, something rare and almost impossible before Hatsune Miku. A lot of amateur musicians have been inspired and motivated by her character image together with her voice and have written songs for her. Many people realized that having a character facilitated writing lyrics for a synthesized singing voice, and that multimodality is an important aspect in singing synthesis.

An important multimodal challenge, therefore, is to generate several attributes of a singer, such as voice, face, and body. The face and body can be realized by computer graphics or robots. An example of simultaneous control of voice and face was shown in the combination of VocaListener [31] and VocaWatcher [50], which imitates singing expressions of the voice and face of a human singer.

In the future, speech synthesis could also be fully integrated with singing synthesis. It will be challenging to develop new voice synthesis systems that could seamlessly generate any voice produced by a human or virtual singer/speaker.

Acknowledgements

The authors would like to thank Alastair Porter for proofreading and Merlijn Blaauw for reviewing the article. Some works presented in the article were supported in part by CREST, JST.

References

[1] X. Rodet, “Synthesis and processing of the singing voice”, in Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA), 2002, pp. 99-108.

[2] P. R. Cook, “Singing voice synthesis: History, current work, and future directions”, Computer Music Journal, 1996, vol. 20, no. 3, pp. 38–46.

[3] J. Sundberg, “The KTH synthesis of singing”, Advances in Cognitive Psychology, 2006, vol. 2, no. 2-3, pp. 131-143.

[4] M. Goto, “Grand challenges in music information research”, in M. Müller, M. Goto, M. Schedl, editors, Multimodal music processing, Dagstuhl Publishing, 2012, vol. 3, pp. 217–225.

[5] A. Kirke and E.R.Miranda, Guide to Computing for Expressive Music Performance, Springer, 2013, Ch. 1.

[6] J. Bonada and X. Serra, “Synthesis of the singing voice by performance sampling and spectral models”, IEEE Signal Processing Magazine, 2007, vol. 24, no. 2, pp. 67–79.

[7] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, B. Banno, “Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown”, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, 2009, pp. 3905 – 3908.

[8] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, “Singing Voice Conversion Method Based on Many-to-Many Eigenvoice Conversion and Training Data Generation Using a Singing-to-Singing Synthesis System”, in APSIPA (Asia-Pacific Signal and Information Processing Association) ASC (Annual Summit and Conference), Dec. 2012, pp. 1-6.

[9] S. Canazza, G. De Poli, C. Drioli, A. Roda, A. Vidolin, “Modeling and control of expressiveness in music performance”, Proc. IEEE Special Issue on Engineering and Music, 2004, vol. 92, no. 4, pp. 686-701.

[10] G. Widmer, “Using AI and machine learning to study expressive music performance: Project survey and first report”, AI Communications, 2001, vol. 14, no. 3, pp. 149-162.

[11] P. N. Juslin, “Five facets of musical expression: A psychologist's perspective on music performance”, Psychology of Music, 2003, vol 31, no. 3, pp. 273-302.

[12] P. N. Juslin and P. Laukka, “Communication of emotions in vocal expression and music performance: Different channels, same code?”, Psychological Bulletin, 2003, vol. 129, no. 5, pp. 770-814.

[13] S. Ternström, “Session on naturalness in synthesized speech and music”, 143rd Acoustical Society of America (ASA) meeting. Available: http://www.pvv.ntnu.no/~farner/sonata/ternstrom02.html, Jun. 2002.

[14] M. Thalén and J. Sundberg, “Describing different styles of singing: A comparison of a female singer’s voice source in ‘classical’, ‘pop’, ‘jazz’ and ‘blues’”, Logopedics Phoniatrics Vocology, 2001, vol. 26, no. 2, pp. 82–93.

[15] A. Friberg, R. Bresin, and J. Sundberg, “Overview of the KTH rule system for musical performance”, Advances in Cognitive Psychology, 2006, vol. 2, no. 2-3, pp. 145–161.

[16] N. Obin, “MeLos: Analysis and Modelling of Speech Prosody and Speaking Style”, Ph.D. dissertation, Université Pierre et Marie Curie-Paris VI, France, 2011. Available: http://articles.ircam.fr/textes/Obin11e/index.pdf.

[17] M. Schröder, “Expressive speech synthesis: Past, present, and possible futures”, Affective information processing, Springer London, May 2009, pp. 111-126.

[18] G. Widmer and W. Goebl, “Computational models of expressive music performance: The state of the art”, Journal of New Music Research, 2004, vol. 33, no. 3, pp. 203–216.

[19] M. Lesaffre, “Music Information Retrieval Conceptual Framework, Annotation and User Behaviour”, Ph.D. dissertation, Ghent University, 2005. Available: https://biblio.ugent.be/publication/3258568.

[20] J. Salamon, E. Gómez, D. P. W. Ellis and G. Richard, "Melody Extraction from Polyphonic Music Signals: Approaches, Applications and Challenges", IEEE Signal Processing Magazine, Mar. 2014, vol 31, no 2, pp. 118-134.

[21] C.J. Plack and A.J Oxenham, “Overview: The present and future of pitch”, in Pitch - Neural Coding and Perception (Springer, New York), Springer Handbook of Auditory Research, 2005, vol. 24, Chap. 1, pp. 1-6.

[22] K. Saino, M. Tachibana, and H. Kenmochi, “A singing style modeling system for singing voice synthesizers”, in Proc. Interspeech, Makuhari, Japan, Sept. 2010, pp. 2894–2897.

[23] J. Sundberg, “The Science of the Singing Voice”, Northern Illinois University Press, 1987.

[24] E. D. Scheirer, “Tempo and beat analysis of acoustic musical signals”, Journal of the Acoustical Society of America, 1998, vol. 103, no. 1, pp. 588-601.

[25] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, “An HMM-based Singing Voice Synthesis System”, in Proc.

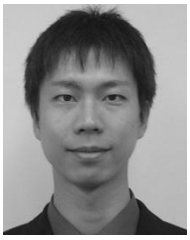
Interspeech, Pittsburgh, PA, USA, Sept. 2006, pp. 1141-1144.
[26] A. Loscos and J. Bonada, "Emulating rough and growl voice in spectral domain", Proc. 7th International Conference on Digital Audio Effects (DAFx), Naples, Italy, Oct. 2004, pp. 49-52.
[27] R. L. de Mantaras and J. Ll. Arcos, "AI and music: From composition to expressive performance", AI magazine, 2002, vol. 23, no. 3, pp. 43-57.
[28] M. Kob, "Singing voice modelling as we know it today", Acta Acustica United with Acustica, 2004, vol. 90, no. 4, pp. 649–661.
[29] Y. Meron, "High Quality Singing Synthesis using the Selection-based Synthesis Scheme", Ph.D. dissertation, University of Tokyo, 1999. Available: http://www.gavo.t.u-tokyo.ac.jp/~meron/sing.html.
[30] J. Janer, J. Bonada, and M. Blaauw, "Performance-driven control for sample-based singing voice synthesis", in Proc. 9th International Conference on Digital Audio Effects (DAFx), 2006, vol. 6, pp. 42–44.
[31] T. Nakano and M. Goto, "VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation", in Proceedings of the 6th Sound and Music Computing Conference (SMC), Jul. 2009, pp. 343-348.
[32] T. Nakano and M. Goto, "VocaListener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics", in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, May 2011, pp.453-456.
[33] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices", in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, Oct. 2007, pp.215-218.
[34] J. Sundberg and J. Bauer-Huppmann, "When does a sung tone start?" in J. Voice, 2007, vol. 21, no. 3, pp. 285-293.
[35] J. Sundberg, "Synthesis of singing, in Musica e Tecnologia: Industria e Cultura per lo Sviluppo del Mezzagiorno", Proceedings of a symposium in Venice, Venedig: Unicopli, 1981, pp. 145-162.
[36] M.C. Marinescu and R. Ramirez, "A machine learning approach to expression modeling for the singing voice", in International Conference on Computer and Computer Intelligence (ICCCI), ASME Press, 2011, vol. 31, no. 12, pp. 311–316.
[37] M. Alonso, "Expressive performance model for a singing voice synthesizer", Master Thesis, Universitat Pompeu Fabra, 2005. Available: http://mtg.upf.edu/node/2223.
[38] R. Bresin and A. Friberg, "Emotional coloring of computer-controlled music performances", Computer Music Journal, Winter 2000, vol. 24, no. 4, pp. 44–63.
[39] H. Kenmochi and H. Ohshita, "VOCALOID - commercial singing synthesizer based on sample concatenation", in Proc. Interspeech, Antwerp, Belgium, Aug. 2007, pp. 4009–4010.
[40] J. Bonada, "Voice processing and synthesis by performance sampling and spectral models", Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, 2008. Available: http://www.mtg.upf.edu/node/1231.
[41] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", in Proc. 6th European Conf. on Speech Communication and Technology (Eurospeech), Budapest, Hungary, 1999, pp. 2347–2350.
[42] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent development of the HMM-based singing voice synthesis system - Sinsy", Proc. Int. Speech Communication Association (ISCA), 7th Speech Synthesis Workshop (SSW7), Tokyo, Japan, Sept. 2010, pp. 211–216.
[43] M. Umbert, J. Bonada, and M. Blaauw, "Generating singing voice expression contours based on unit selection", Stockholm Music Acoustics Conference (SMAC), Stockholm, Sweden, Aug. 2013, pp. 315–320.
[44] M. Umbert, J. Bonada, and M. Blaauw. "Systematic Database Creation for Expressive Singing Voice Synthesis Control", Proc. Int. Speech Communication Association (ISCA), 8th Speech Synthesis Workshop (SSW8), Barcelona, Spain, Sept. 2013, pp. 213-216.
[45] D. Campbell, E. Jones, and M. Glavin, "Audio quality assessment techniques - A review, and recent developments", Signal Processing, Aug. 2009, vol. 89, no. 8, pp. 1489-1500.
[46] M. Chu, and H. Peng, "An objective measure for estimating MOS of synthesized speech", Proc. 7th European Conf. on Speech Communication and Technology (Eurospeech), Aalborg, Denmark, Sept. 2001 pp. 2087-2090.
[47] S. Möller, F. Hinterleitner, T. H. Falk, and T. Polzehl, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems", in Proc. Interspeech, Makuhari, Japan, Sept. 2010, pp. 1325-1328.
[48] H. Katayose, M. Hashida, G. De Poli, and K. Hirata, "On Evaluating Systems for Generating Expressive Music Performance: the Rencon Experience", Journal of New Music Research, 2012, vol. 41, no. 4, pp. 299-310.
[49] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]", IEEE Robotics & Automation Magazine, June 2012, vol. 19, no. 2, pp. 98-100.
[50] M. Goto, T. Nakano, S. Kajita, Y. Matsusaka, S. Nakaoka, and K. Yokoi, "VocaListener and VocaWatcher: Imitating a Human Singer by Using Signal Processing", in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, March 2012, pp. 5393-5396.

Authors

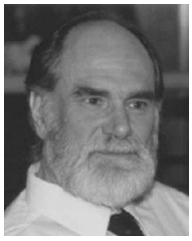
Martí Umbert (marti.umbert@upf.edu) is currently a Ph.D. candidate in the Music Technology Group at the Universitat Pompeu Fabra (UPF) in Barcelona, Spain. After obtaining his degree in Telecommunications at the Universitat Politècnica de Catalunya in 2004 he worked on speech technologies both at the university and at the private sector. In 2010 he obtained the MSc in Sound and Music Computing from the UPF. His research is carried out within the Audio Signal Processing team and he is interested in singing voice synthesis and expression modeling based on unit selection, and corpus generation.

Jordi Bonada (jordi.bonada@upf.edu) Jordi Bonada received the Ph.D. degree in Computer Science and Digital Communications from Universitat Pompeu Fabra (UPF), Barcelona, Spain, in 2009. He is currently a senior researcher at the Music Technology Group from UPF. He is mostly interested in singing voice modeling and synthesis. His research trajectory has an important focus on technology transfer acknowledged by more than 50 patents. Dr. Bonada has received several awards including the Rosina Ribalta prize by Epson Foundation in 2007 and the Japan Patent Attorneys Association Prize by the Japan Institute of Invention and Innovation in 2013.

Masataka Goto (m.goto@aist.go.jp) received the Doctor of Engineering degree from Waseda University in 1998. He is currently a Prime Senior Researcher and the Leader of the Media Interaction Group at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. Over the past 23 years, he has published more than 220 papers in refereed journals and international conferences and has received 40 awards, including several best paper awards, best presentation awards, the Tenth Japan Academy Medal, and the Tenth JSPS PRIZE. In 2011, as the Research Director he began OngaCREST Project, a 5-year JST-funded research project (CREST) on music technologies.



Tomoyasu Nakano (t.nakano@aist.go.jp) received the Ph.D. degree in informatics from University of Tsukuba, Tsukuba, Japan in 2008. He is currently working as a Senior Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. His research interests include singing information processing, human-computer interaction, and music information retrieval. Dr. Nakano has received several awards including the IPSJ Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) and the Best Paper Award from the Sound and Music Computing Conference 2013 (SMC2013). He is a member of the IPSJ and the Acoustical Society of Japan (ASJ).



Johan Sundberg (jsu@csc.kth.se) born 1936, Ph.D. musicology, doctor honoris causae 1996 University of York, UK and 2014 Athens University. He had a personal chair in Music Acoustics at KTH and founded and its music acoustics research group until 2001. His research concerns particularly the singing voice and music performance. He has written The Science of the Singing Voice (1987) and The Science of Musical Sounds (1991). He is member of the Royal Swedish Academy of Music, the Swedish Acoustical Society (President 1976-81), and fellow of the Acoustical Society of America, Silver Medal in Musical Acoustics 2003.