Freixes *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2019) 2019:22
https://doi.org/10.1186/s13636-019-0163-y

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

A unit selection text-to-speech-and-singing synthesis framework from neutral speech: proof of concept

Marc Freixes*, Francesc Alías and Joan Claudi Socoró

Abstract

Text-to-speech (TTS) synthesis systems have been widely used in general-purpose applications based on the generation of speech. Nonetheless, there are some domains, such as storytelling or voice output aid devices, which may also require singing. To enable a corpus-based TTS system to sing, a supplementary singing database should be recorded. This solution, however, might be too costly for eventual singing needs, or even unfeasible if the original speaker is unavailable or unable to sing properly. This work introduces a unit selection-based text-to-speech-and-singing (US-TTS&S) synthesis framework, which integrates speech-to-singing (STS) conversion to enable the generation of both speech and singing from an input text and a score, respectively, using the same neutral speech corpus. The viability of the proposal is evaluated considering three vocal ranges and two tempos on a proof-of-concept implementation using a 2.6-h Spanish neutral speech corpus. The experiments show that challenging STS transformation factors are required to sing beyond the corpus vocal range and/or with notes longer than 150 ms. While score-driven US configurations allow the reduction of pitch-scale factors, time-scale factors are not reduced due to the short length of the spoken vowels. Moreover, in the MUSHRA test, text-driven and score-driven US configurations obtain similar naturalness rates of around 40 for all the analysed scenarios. Although these naturalness scores are far from those of vocaloid, the singing scores of around 60 which were obtained validate that the framework could reasonably address eventual singing needs.

Keywords: Text-to-speech, Unit selection, Speech synthesis, Singing synthesis, Speech-to-singing

1 Introduction

Text-to-speech (TTS) synthesis systems have been widely used to generate speech in several general-purpose applications, such as call-centre automation, reading emails or news, or providing travel directions, among others [1]. However, there are other domains that may require the eventual generation of singing in addition to speech. For instance, in storytelling [2, 3], when one of the characters sings at one point in the story, or in voice output communication aid devices for individuals with vocal disabilities [4] to allow them not only to talk, but also to sing. Moreover, a TTS with singing capabilities could also be useful in assistive technologies, where the incorporation of songs has been proved to be an effective form of improving the engagement of autistic children [5], or to reduce the procedural distress in children with cancer [6], or to augment the positive memories of people with dementia [7], to name a few.

In this sense, it is worth mentioning that early works on speech synthesis already enabled the generation of both speech and singing (e.g. see [8]), as they stood on a source-filter model inspired by the classical acoustic theory of voice production [1]. However, the difficulty of defining and adjusting the necessary control parameters to obtain high-quality speech led the research towards data-driven approaches [1]. Although some approaches used diphone-based TTS systems to generate singing [9, 10], most works opted to use databases specifically recorded for singing purposes [11–13]. Meanwhile, the speech synthesis investigations also moved to corpus-based approaches, deploying TTS systems based on unit selection (US), hidden Markov models (HMM) or hybrid approaches, and more

*Correspondence: marc.freixes@salle.url.edu
Grup de recerca en Tecnologies Mèdia (GTM), La Salle - Universitat Ramon Llull, Quatre camins, 30, 08022 Barcelona, Spain

recently, including deep neural networks (DNN) (e.g., [14, 15]). Even though these systems can deliver very natural synthetic speech [16], as far as we know, they are not able to speak and sing at the same time.

In order to add singing capabilities to a corpus-based TTS system, the first idea that may come to mind is to incorporate a supplementary singing database. However, occasional singing needs do not justify the cost of building an additional corpus, which may become unfeasible if the original speaker is unavailable or unable to sing properly [17, 18]. As an alternative, we could take advantage of those approaches which focus on the production of singing from speech following the so-called speech-to-singing (STS) conversion [19–21]. These techniques can be applied to the output of a TTS system to transform speech to singing by maintaining the identity of the speaker [18, 22]. However, this straightforward approach has been proved suboptimal in terms of flexibility and computational costs [18].

Building on the preliminary approach presented in [18], this work introduces a unit selection-based text-to-speech-and-singing (US-TTS&S) synthesis framework that allows the generation of both speech and singing from an input text and a score, respectively, using the same neutral speech corpus. To this end, the framework incorporates speech-to-singing (STS) conversion within a TTS system pipeline. The viability of the proposal is evaluated objectively and subjectively through a proof-of-concept implementation of the US-TTS&S framework using a 2.6-h Spanish neutral speech corpus.

The paper is structured as follows. Section 2 reviews the singing and speech-to-singing literature. Then, Section 3 describes the proposed US-TTS&S framework and the proof-of-concept implementation. The methodology employed for the objective and the subjective evaluation is detailed in Section 4. Finally, after presenting and discussing the results (Section 5), the conclusions of this work are drawn in Section 6.

2 Related work

This section includes a review of the singing synthesis approaches which are closely related to speech synthesis and a description of speech-to-singing techniques.

2.1 Singing synthesis

Until the late 1980s, most of the singing synthesis approaches were closely linked to sound synthesis [23] or to speech synthesis (see [24] and references therein). The latter correspond to first generation synthesis systems, where according to a synthesis specification (verbal component, pitch values, and durations), a rule-based control drives a source-filter model built on the classical acoustic theory of voice production. On the one hand, articulatory speech synthesis [25] was used to generate one of the first synthetic singing examples[1]. This technology evolved giving rise to systems such as SPAM/Singer [8], which could be used for TTS and singing synthesis through control files [26]. On the other hand, formant speech synthesis inspired the development of singing systems as the MUSSE (MUsic and Singing Synthesis Equipment) and the subsequent MUSSE DIG (MUSSE, DIGital version) [27] or the CHANT project [28]. First-generation rule-based systems gave way to data-driven approaches mainly due to the difficulty of generating the control parameters to get high-quality results [1]. However, formant synthesis is still used nowadays in the context of performative singing synthesis [29], where flexibility and real time are the main issues.

In second-generation synthesis systems, a unit (typically a diphone) for each unique type was recorded. Pitch and timing of units were modified applying signal processing techniques to match the synthesis specification [1]. Some works exploited signal processing capabilities to generate singing from a spoken database. Flinger [9] for instance used residual LPC synthesis and provided several modules in order to enable the Festival TTS system [30] to sing. MBROLA was also used to generate both speech and singing from speech units [10, 31]. Similarly, the Ramcess synthesiser [32] generated singing by convolving vocal tract impulse responses from a database with an interactive model of the glottal source. However, the data-driven paradigm of second generation synthesis systems naturally led to the creation of singing databases.

Finally, it should be noted that there have been some recent attempts to produce singing from speech in a corpus-based TTS system. Some works used the system to get a spoken version of the song and transform it into singing by incorporating a signal processing stage. For instance, in [22], the synthetic speech was converted into singing according to a MIDI file input, using STRAIGHT to perform the analysis, transformation and synthesis. In [17], an HMM-based TTS synthesiser for Basque was used to generate a singing voice. The parameters provided by the TTS system for the spoken version of the lyrics were modified to adapt them to the requirements of the score.

2.2 Speech-to-singing

Speech-to-singing conversion is the task of transforming the spoken lyrics of a song into singing, while retaining the identity of the speaker and the linguistic content [33]. In [20], the authors proposed a method to transform speech into singing, by modifying the pitch contour, the duration of the phonemes and the spectrum according to the analysis of the features of the singing voice. Phoneme target durations were obtained by applying STS duration conversion rules derived from the analysis of real

[1] https://ccrma.stanford.edu/~jos/wav/daisy-klm.wav

performances. The pitch contour was derived from a stepwise melody curve by applying a filtering that models the behaviour and dynamics of the fundamental frequency ($F0$) in singing: preparation, overshoot, fine fluctuation, and vibrato. Finally, two spectral control models were applied to the envelope to add the singing formant and to apply a formant amplitude modulation that was synchronised with the vibrato. Analysis, transformation, and synthesis were carried out using STRAIGHT [34].

In order to obtain more natural contours, other approaches have used real singing performances, but they require spoken and sung parallel recordings. In [19], a set of phrases was recorded by a female singer to get a spectral envelope database. The same speech sentences, recorded by an actor, were time-stretched, transposed, and aligned with the singing phrases. Finally, the spectral envelope from the singer database was transferred to the speech signal. The transformation was performed by a phase vocoder in this case. However, improved signal models were subsequently proposed [35, 36]. In [21], I^{2}R Speech2Singing system was presented. This application was able to convert speech or poor singing into high-quality singing, using a template-based conversion [37] with professional singing as a reference model. Parallel singing templates were aligned with the speech input in a 2-step dynamic time warping-based method. Thus, the pitch contour could be derived from actual singing voice and applied to the input speech through STRAIGHT. An improved dual alignment scheme for this system has been recently presented in [38].

Finally, apart from appropriate timing and F0 contours, spectral transformation is a very important issue in speech-to-singing conversion. Voice conversion and model adaptation techniques were extended to this scenario in [39], using a large collection of singing recordings and their corresponding spoken lyrics. The comparison between these methods and the spectral transformation applied in [20] showed that model adaptation outperforms the other approaches in singing quality and similarity provided there is enough data.

3 US-TTS&S synthesis framework from neutral speech

This section is organised as follows. Section 3.1 describes the proposed US-TTS&S synthesis framework. Next, Section 3.2 details the proof-of-concept implementation of the framework.

3.1 Framework

The block diagram of the proposed synthesis framework is depicted in Fig. 1. It consists of two main subsystems: the text-to-speech subsystem (at the top), which allows the framework to produce neutral synthetic speech for a given input text, and the speech-to-singing subsystem (at the bottom), which provides the framework with singing capabilities.

In the speech mode, the input text is analysed by the Natural Language Processing (NLP) module, which yields a linguistic target (including the phonetic transcription and the linguistic context) and predicts a speech prosodic target (i.e. phrasing and intonation appropriate for the message). The unit selection block searches the corpus for the units that best meet these targets and that can be smoothly joined. Finally, the parametric representations of the selected units are concatenated, thus obtaining a stream of speech parameters that is rendered into synthetic speech through the waveform generation module.

In the singing mode, the input score S, which contains the lyrics as syllables assigned to the notes, is parsed by the score processing module, which extracts the lyrics, integrates score and phonetic information, and provides a singing prosodic target to perform the unit selection according to S and the optional tempo and transposition values. Subsequently, the transformation module converts the retrieved speech parameters into the singing ones, according to the controls computed by the expression control generation module. Finally, the waveform generation module renders the modified parameters into synthetic singing.

The following subsections describe the key modules for the singing mode.

3.1.1 Score processing

This module joins the syllables extracted from S in order to get the full lyrics of the song, which are fed into the TTS subsystem (see Fig. 1). Subsequently, it obtains the links between the phonetic transcription of the lyrics (provided by the NLP module) and the notes. To this end, the phonemes are distributed among the notes according to the assignment of syllables to notes in the score. Furthermore, since a note onset coincides with a vowel [40], the preceding consonants are assigned to the preceding note.

Moreover, this module allows for score transposition according to the input value x, obtaining thereby a transposed score Sx. This score is then used, together with phoneme-note links and tempo T, to compute the singing prosodic target (see Section 3.1.2) and to generate the expression controls (see Section 3.1.3).

With regard to tempo, the value of T in beats per minute (bpm) can be extracted from S, or alternatively indicated as an input of the synthesis framework (see Fig. 1). Tempo is used to compute the duration of each note according to its note value (e.g., quarter note, eighth note).

Regarding score transposition, this process consists of moving the entire set of notes up or down in pitch by a constant interval in order to fit it within the vocal range of the singer. Accordingly, the notes of S are shifted to get the score Sx, whose pitch range midpoint $F0_m^{\mathrm{Sx}}$ is x semitones

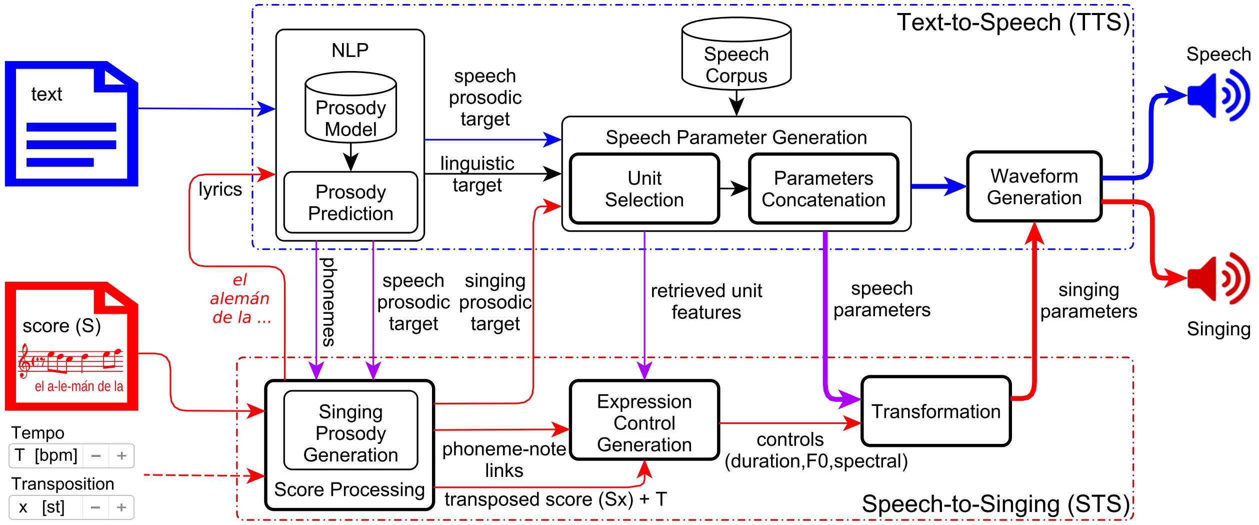


Fig. 1 US-TTS&S framework. Block diagram of the unit-selection text-to-speech and singing (US-TTS&S) synthesis framework from neutral speech. In the speech mode, an input text is converted into synthetic speech by the TTS subsystem (above in the blue box). In the singing mode, the incorporation of the speech-to-singing (STS) subsystem (below in the red box) enables the framework to produce synthetic singing from an input score S (containing both the notes and the lyrics), considering optional input values: tempo *T* in beats per minute and transposition *x* in semitones

above the speech corpus vocal range midpoint $F0_m^{\mathrm{C}}$, which represents an intermediate value within the pitch range covered by the vowels in the corpus C. To this end, the note pitches in S are translated into an integer notation following a 12-tone equal temperament, which divides the octave into 12 semitone steps equally spaced on a logarithmic scale. Thus, a note number $N^{\mathrm{S}}(i)$ is obtained for each note in S, where $i = \{1..K\}$, being K the total number of notes in the score S. Subsequently, the note numbers for Sx are computed as

$$N^{\mathrm{Sx}}(i) = N^{\mathrm{S}}(i) + x - d\left(F0_m^{\mathrm{C}}, F0_m^{\mathrm{S}}\right) \tag{1}$$

where

$$d\left(F0_m^{\mathrm{C}}, F0_m^{\mathrm{S}}\right) = \left[12log_2\left(\frac{F0_m^{\mathrm{S}}}{F0_m^{\mathrm{C}}}\right)\right] \tag{2}$$

is the distance in semitones from the speech corpus vocal range midpoint $F0_m^{\mathrm{C}}$ to the input score pitch range midpoint $F0_m^{\mathrm{S}}$, and $[\cdot]$ denotes that the result of the operation is rounded to the nearest integer. Since the perception of pitch is logarithmic, $F0_m^{\mathrm{S}}$ is computed from the lowest and the highest note as the geometric mean of their $F0$ values, i.e.

$$F0_m^{\mathrm{S}} = \sqrt{F0_{\mathrm{min}}^{\mathrm{S}} \cdot F0_{\mathrm{max}}^{\mathrm{S}}}. \tag{3}$$

3.1.2 Singing prosody generation

This block translates the note durations and $F0$s obtained from Sx and T into a prosodic representation of the singing target consisting of phonetic timing and $F0$s. This singing prosodic target enables the US-TTS&S framework to perform the unit selection according to Sx and T. The phonetic timing is obtained by adjusting the duration of the phonemes so that they fit the duration of the notes to which they are linked. Similarly, the $F0$ of each note is assigned to its phonemes considering that the note $F0$ is reached at the vowel onset, so the transition occurs in the precedent phoneme [40].

3.1.3 Expression control generation

Expression control in singing synthesis, also known as performance modelling, consists in the manipulation of a set of voice features (e.g., phonetic timing, pitch contour, vibrato, timbre) that relates to a particular emotion, style, or singer [41]. Accordingly, the expression control generation module provides the duration, $F0$, and spectral controls required by the transformation module to convert the sequence of speech parameters into singing parameters. To this end, and following the phoneme-note links, this module aligns the units retrieved by the US block with the notes, and generates the controls to transform the spoken features (durations, $F0$ and spectra) into singing ones in accordance with Sx and T. Since obtaining control parameters that are perceived as natural is one of the main issues regarding singing synthesis, several approaches can be found in the literature (see [41] and references therein for further details).

3.1.4 Speech parameter generation and transformation

In contrast to *pure* unit selection, where an overlap and add (OLA) method is applied to the retrieved units, with the aim of modifying the original waveforms as little as possible [1], the US-TTS&S framework is based on a parametric representation of the speech. This enables the use of more flexible processing techniques to address the highly significant transformations (including spectral ones) involved in the STS conversion.

The framework signal processing pipeline consists of three modules. The speech parameter generation module performs the unit selection (according to the linguistic and prosodic targets) and concatenates the parametric representation of the selected units to obtain a speech parameter sequence. In the speech mode, this sequence is directly fed into the waveform generation module to produce synthetic speech. Conversely, in the singing mode, the sequence is previously processed by the transformation module, which applies time-scaling, pitch-scaling, and spectral transformations to convert the speech parameters into singing ones.

3.2 Proof-of-concept implementation

In the following paragraphs, the main elements of the implementation of the US-TTS&S framework are described.

3.2.1 Text-to-speech subsystem

The US-TTS system of La Salle-Universitat Ramon Llull [42] has been used as text-to-speech subsystem. This TTS synthesis system includes a case-based reasoning (CBR) prosody prediction block, trained with acoustic prosodic patterns from the speech corpus, and a unit selection block following a classical scheme [43]. This block retrieves the units that minimise the prosodic, linguistic, and concatenation costs (see [42] for more details). The weights for the prosodic target and concatenation subcosts were perceptually tuned by means of active interactive genetic algorithms for speech synthesis purposes [44].

The Time-Domain Pitch Synchronous Overlap and Add (TD-PSOLA) waveform generation used in the original US-TTS system has been replaced by a harmonic plus noise model (HNM) implementation [45]. Accordingly, the corpus has been parameterised with HNM representation. The harmonic component (for the voiced frames) is modelled by a sum of sinusoids (each with a certain amplitude and phase) at the multiples of the fundamental frequency up to the 5 kHz maximum voiced frequency [46]. This component is subtracted from the speech signal to get the stochastic (noise) component, which is analysed using an autoregressive model and it is represented with 15-order linear prediction coefficients (LPC) and the noise variance [46]. The HNM analysis has been performed pitch-synchronously, applying a window around the centre of gravity to avoid phase mismatches when units are concatenated [47].

3.2.2 Score processing

The proof-of-concept implementation of this module has adopted the MusicXML[2] format for the score S. To this end, the scripts from Nichols et al. [48] have been considered. In MusicXML, each syllable of the lyrics is assigned to a note with the *lyric* element. This contains a *text* element with the syllable and a *syllabic* element that indicates how the syllable fits into the word. The latter can take the values *single*, *begin*, *end*, or *middle*, and is used to recompose the words and obtain the whole text of the song. The syllabic element also provides the syllabic distribution, which is considered to assign the phonemes from each word to their corresponding notes. An example of this alignment is depicted at the top of Fig. 2.

With regard to the $F0$, each MusicXML note in S is parsed into a MIDI note number $N^{\mathrm{S}}(i)$, whose $F0$ is computed as

$$F0^{\mathrm{S}}(i) = 440 \cdot 2^{(N^{\mathrm{S}}(i)-69)/12}, \tag{4}$$

since the MIDI note 69 corresponds to A4 (440 Hz)[3]. If a transposition value of x semitones is introduced into the framework, the shifted MIDI note numbers for Sx are computed following Eq. (1), (2), and (3).

The speech corpus vocal range is defined from the $F0$ mean values of the vowels within it. According to this, the speech corpus vocal range midpoint is computed in this implementation as

$$F0_m^{\mathrm{C}} = \sqrt{F0_5^{\mathrm{C}} \cdot F0_{95}^{\mathrm{C}}}, \tag{5}$$

where $F0_5^{\mathrm{C}}$ and $F0_{95}^{\mathrm{C}}$ are the 5th and the 95th corpus vowel $F0$ percentiles, respectively, thus avoiding possible outliers.

3.2.3 Singing prosody generation

This block generates a singing prosodic target according to the durations and $F0$s obtained from score Sx and tempo T.

On the one hand, the phoneme durations predicted by the prosodic model (represented below the score S in Fig. 2) are adjusted to fit the note durations by applying the STS conversion rules derived by Saitou from the analysis of real performances [20]:

1. When phonemes tied to a note have to be shortened, their original durations are multiplied by the same factor.
2. When phonemes have to be stretched, three parts are differentiated around the boundary between a consonant and a vowel: the consonant, the transition (from 10 ms before to 30 ms after the boundary), and the vowel.

[2] https://www.musicxml.com

[3] https://www.midi.org/specifications

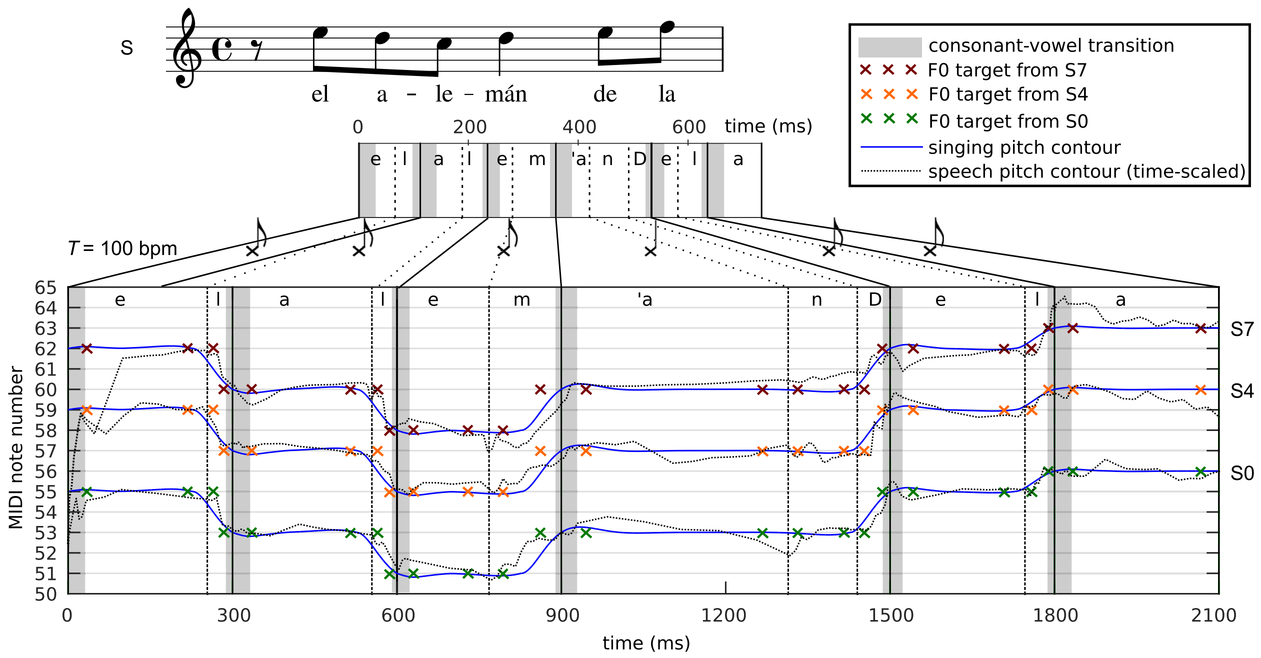


Fig. 2 Example of a song excerpt synthesised with transposed scores S0, S4, and S7. The phonemes from the lyrics phonetic transcription are represented below the input score S, together with their durations, which are (i) predicted from the lyrics by the NLP module when computing the singing prosodic target for the US block (see Fig. 1), or; (ii) those of the retrieved speech units when generating the expression controls. At the bottom, the phoneme durations have been time-scaled to fit the note durations. The crosses represent the *F*0 values of the singing prosodic targets obtained from S0, S4, and S7. The pitch contours (time-scaled) of the retrieved speech units are depicted as dashed grey lines. Finally, the solid blue lines represent the singing pitch contours generated by the expression control generation module. The score-driven US configuration Sxp and $T = 100$ bpm have been used for this example

(a) The consonant part is extended according to fixed rates (1.58 for a fricative, 1.13 for a plosive, 2.07 for a semivowel, 1.77 for a nasal, and 1.13 for a /y/).

(b) The transition part (depicted as a shadowed area in Fig. 2) is not extended.

(c) The vowel part is extended until the phoneme fits the note duration.

In the current implementation the transition length within the vowel (30 ms) has been limited to a maximum of half of its duration, since the corpus contains very short vowels.

On the other hand, the *F*0 target (represented by crosses in Fig. 2) is assigned at a semiphoneme level. The *F*0 from each note in Sx is assigned to all its corresponding semiphonemes, except in the transitions where the right semiphoneme receives the *F*0 of the following note.

3.2.4 Expression control generation

This module computes the duration, *F*0, and spectral controls required to perform the STS conversion, in accordance with Sx and *T*.

Regarding the duration control, the durations of the phonemes retrieved by the US block are scaled to fit the durations of the notes by applying the conversion rules detailed in Section 3.2.3. The correspondence between the original and the scaled durations will drive the time-scaling process performed by the transformation module.

With respect to the *F*0 control, a singing pitch contour (the blue solid lines in Fig. 2) is obtained following the approach described in [20]. According to this, a stepwise pitch contour is built from *F*0s and durations of the notes. Then, this contour is filtered to obtain the singing *F*0 characteristic fluctuations: overshoot, preparation, and fine fluctuation. Figure 3 depicts an example of a pitch curve generation. Overshoot (upper right) is obtained by directly filtering the stepwise contour. Alternatively, preparation (upper left) can be obtained by filtering (from the end towards the beginning) a slightly delayed version of the stepwise curve. The mix (bottom) of both fluctuations is obtained by applying the masks (middle), which prioritise the overshoot at the beginning of the note, preparation at the end, and consider a simple cross-fading in between. In this proof of concept, the implementation of vibrato is left for future research.

Finally, the spectral control tries to emulate the singing formant by emphasising the spectral envelope peak around 3 kHz within the vowels [20].

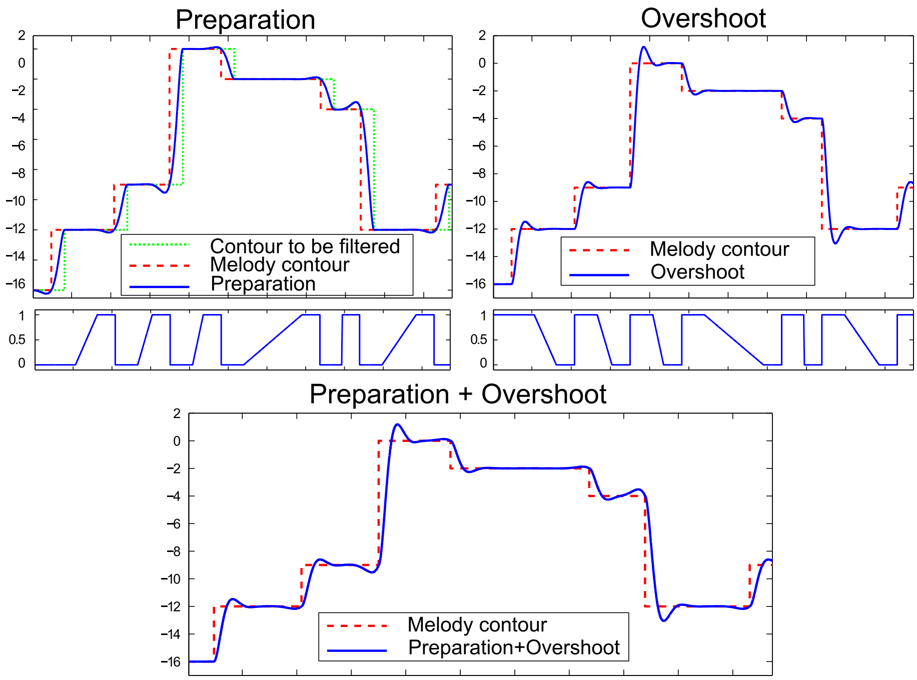


Fig. 3 Singing pitch curve generation. Preparation (upper left), overshoot (upper right), the applied masks (middle), and the resulting mix (bottom)

3.2.5 Speech parameter generation and transformation

The HNM parameters of the retrieved units are concatenated, removing pitch and spectrum mismatches by applying a simple linear interpolation technique around the joins [47]. Transformation and synthesis are performed pitch-synchronously. Thus, when a prosody modification is performed, the HNM parameters in the new epochs are obtained pitch-synchronously through the time-domain linear interpolation of the original parameters. Furthermore, if pitch scaling is done, amplitudes and phases are interpolated in frequency to preserve the original spectral envelope shape [49]. The new harmonic amplitudes are calculated by the linear interpolation of the spectral envelope in a logarithmic amplitude scale. The phases of the target harmonics are obtained by interpolating the real and the imaginary parts of the harmonic complex amplitudes at the new frequencies. Finally, the amplitudes are scaled to preserve the energy despite the variation in the number of harmonics.

4 Methods

This section describes the methods used for the evaluation of the proposed US-TTS&S synthesis framework through the proof-of-concept implementation using a Spanish corpus. The study has been carried out for three vocal ranges and two tempos, and considering a text-driven and three score-driven US configurations. The experiments setup is described in Section 4.1. Then, the objective evaluation (Section 4.2) analyses the magnitude of the transformations required by the STS process to allow the framework to sing. Finally, the subjective evaluation (Section 4.3) assesses both the singing capabilities of the framework together with the naturalness of the synthesised singing.

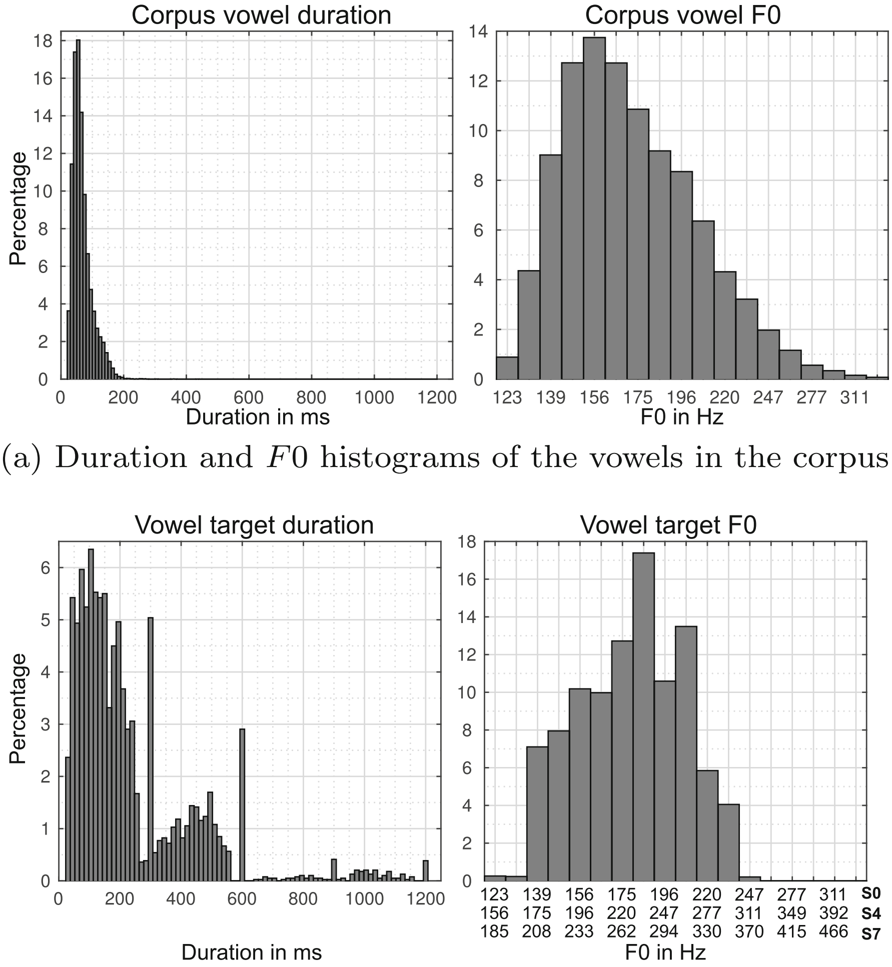
4.1 Experiments setup

4.1.1 Corpus

The experiments have been performed using a 2.6-h Spanish neutral speech corpus recorded by a female professional speaker [50]. The duration and $F0$ histograms of the corpus vowels are depicted in the Fig. 4a. Regarding duration, about half of the vowels last 50 ms or less, and there are virtually none beyond 200 ms. The $F0$ histogram has been depicted so each bin coincides with a semitone in an equal temperament. Even though the corpus contains vowels from 123 until 330, 3 out of 4 are between 139 and 196, so only cover 7 semitones. The 5th and the 95th percentiles ($F0_5^C$ and $F0_{95}^C$) are 134.4 Hz and 235.5 Hz, respectively. Therefore, the corpus vocal range midpoint is $F0_m^C = \sqrt{F0_5^C \cdot F0_{95}^C} = 178$ Hz.

4.1.2 Vocal ranges and tempos

The first evaluation scenario considered corresponds to singing in the corpus vocal range (S0). However, in order to evaluate the capability of the proposed US-TTS&S system to work in a singer vocal range, a contralto set up has been also examined; this is 7 semitones above the speech corpus pitch range midpoint (S7). Moreover, the study has



(a) Duration and $F0$ histograms of the vowels in the corpus.

(b) Vowel target duration and $F0$ histograms predicted from the test score dataset sung with $T = 100$ bpm, and 0, 4 and 7 semitones above the speech corpus vocal range midpoint S0, S4 and S7, respectively).

Fig. 4 Corpus (**a**) and target (**b**) vowel duration and *F*0 distributions

been completed with an intermediate anchor point (S4). Finally, regarding the tempo, two values have been considered: $T = 100$ bpm corresponding to a moderate speed, and a slow one ($T = 50$ bpm).

4.1.3 Unit selection configurations

The evaluation has included a text-driven US configuration, MLC, which considers linguistic (L) and concatenation (C) costs, and the prosodic target predicted from the lyrics by the CBR prosodic model (M). This would correspond to the default US-TTS setting.

Moreover, the study has also considered three score-driven configurations. In this case, the prosodic target is that obtained by the singing prosody generation block according to Sx and T. These configurations are SxpdLC, which uses the pitch (p) and duration (d) from the score instead of those from the model, SxpdC, which also disables the linguistic cost, and finally Sxp, which only considers the pitch.

4.2 Objective evaluation

The objective analysis has been conducted by feeding a score dataset into the framework to be sung in the aforementioned vocal ranges and tempos with the considered US configurations. Then, the pitch and time-scale factors required to transform the retrieved units into singing have been computed. More specifically, the analysis has been focused on the vowels, where the bulk of the signal processing takes place. Moreover, the approach described in [51] has been implemented to get a binary concatenation quality prediction (poor/good) for each join (within the vowels). The subsequent paragraphs describe the details of the experiments.

4.2.1 Score test dataset

From a score compilation of songs for children [52], a subset of 279 musical phrases has been selected, by applying a greedy algorithm [53] to ensure its phonetic coverage in Spanish. This has resulted in a dataset containing 3899 notes, which spans 29 min and 57 s with $T = 100$ bpm and 59 min and 54 s for $T = 50$ bpm.

Figure 4b presents the vowel duration and $F0$ targets generated from the dataset by the singing prosody generation block. The left of Fig. 4b shows the histogram of the vowel duration target for the score dataset sung at 100 bpm, while the right section depicts the histogram of the vowel $F0$ target for the dataset performed with S0, S4, and S7.

4.2.2 Transformation requirements

A time-scale factor (β) has been calculated for each retrieved vowel as

$$\beta = \frac{\mathrm{Dur}_{\mathrm{tgt}} - \mathrm{Dur}_{\mathrm{trn}}}{\mathrm{Dur}_{\mathrm{orig}} - \mathrm{Dur}_{\mathrm{trn}}} \qquad (6)$$

where $\mathrm{Dur}_{\mathrm{tgt}}$ is the singing target duration and $\mathrm{Dur}_{\mathrm{orig}}$ is the original duration of the retrieved vowel. When the vowel is stretched, Dur_{trn} accounts for the duration of the unscaled transition (shadowed areas in Fig. 2), otherwise $\mathrm{Dur}_{\mathrm{trn}} = 0$.

Regarding the pitch-scale factors (α), since the core US-TTS works with diphones, we have obtained two values for each vowel, i.e., one for each semiphoneme. The pitch-scale factor has been computed as

$$\alpha = \frac{F0_{\mathrm{tgt}}}{\mathrm{mean}(F0_{\mathrm{orig}})} \qquad (7)$$

where $F0_{\mathrm{tgt}}$ is the target $F0$ assigned from Sx, and $\mathrm{mean}(F0_{\mathrm{orig}})$ is the mean of the $F0$ values within the retrieved semiphoneme. Pitch-scale factors are expressed in number of semitones as $\alpha_{st} = 12\log_2(\alpha)$, since these units are more meaningful from a musical point of view and closer to the logarithmic perception of the pitch.

Moreover, transformation factors have been categorised taking into account reference values in the literature. Regarding time-scale factors, authors in [54] considered the values below 4 as moderate, whereas in [55] only factors smaller than 2.5 received this consideration. According to this, time-scale factors have been grouped in three categories: low (< 2.5), moderate (2.5, 4], and high (> 4). Similarly, pitch-scale values have also been categorised according to typical values [55] (see Table 1).

Finally, the statistical significance of the differences among the results has been analysed using the Wilcoxon signed-rank test for the transformation factors, and McNemar for the discretised factors .

Table 1 Pitch-scale intervals expressed in absolute number of semitones ($|\alpha_{st}|$) and as multiplying factors (α)

$\lvert\alpha_{st}\rvert$	[0–4]	(4–7]	(7–12]	> 12
$\alpha < 1$	[0.8–1)	[0.67–0.8)	[0.5–0.67)	< 0.5
$\alpha > 1$	(1–1.26]	(1.26–1.50]	(1.50–2]	> 2

4.3 Subjective evaluation

4.3.1 MUSHRA test setup

The subjective evaluation is based on the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test [56], and it was done using the Web Audio Evaluation Tool [57]. For the evaluation, five sentences were chosen from the speech corpus so that their phonetic stress distribution could coincide with the music stressed beats. These sentences were set to music using eighth notes (the most common note value), thus getting five scores. These songs were synthesised in the 3 vocal ranges (S0, S4, and S7) and the 2 tempos (100 bpm and 50 bpm) considering the 4 US configurations under study. The obtained audios were analysed following the procedure described in Section 4.2 to check that the transformation factors obtained for the different US configurations fit with those seen in the objective evaluation with the score dataset. The audios generated for one of the five scores have been provided as Additional files 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, and 24.

Forty-nine Spanish native speakers took part in the test. From the 30 evaluation sets (5 scores × 3 vocal ranges × 2 tempos), each user evaluated 6 sets corresponding to the 6 case scenarios (3 vocal ranges × 2 tempos). For each set, the participants were told to rate different versions of the same melody compared to a reference on a scale of 0 to 100. Specifically, they were told to evaluate the naturalness and the singing (i.e. how well sung is each stimuli regardless the naturalness). Moreover, they were instructed to give the highest score to the reference. Thus we excluded 14.5% of the sets where participants rated the hidden reference below 70.

Regarding the singing evaluation, the score performed by Vocaloid [58] was used as the upper reference and the lyrics synthesised by the TTS (i.e. not sung) as the lower anchor (see, for example, Additional file 25). Since the STS process applied is the same for all the US configurations, only MLC was included together with the hidden reference and the anchor to minimise the fatigue of the participants. For the naturalness assessment, the upper reference was the original sentence from the corpus, i.e. natural speech (see, for example, Additional file 26), while no lower anchor was available. In this case, 7 stimuli were evaluated within each set: MLC, the 3 score-driven configurations (Sxp, SxpdC, and SxpdLC), Vocaloid (V), and the hidden reference.

5 Results and discussion

This section presents and discusses the results obtained from both the objective and the subjective evaluation.

5.1 Objective evaluation

5.1.1 Pitch-scale and concatenation analysis

The distributions of the pitch-scale factors (α_{st}) required to convert the retrieved spoken units into singing are depicted in Fig. 5. Their probability densities are represented by violinplots superposed on the standard boxplots, whose whiskers are set to 2nd and 98th percentiles. The percentages for the categorised pitch-scale factors and for concatenation quality can be seen in Table 2. The values obtained at the two tempos have been included for the configurations that consider durations from the score (SxpdC and SxpdLC). However, since the differences due to the tempo are very small, only the distributions obtained with $T = 100$ bpm have been depicted in Fig. 5.

When singing in the corpus vocal range (look at S0 scenario in Fig. 5), the distribution of pitch-scale factors is centred around 0 semitones in all the configurations. The interval defined by the 2nd and 98th percentiles ranges from [− 6.9,8.2] for MLC to [− 0.8,0.9] for S0p. Therefore, the distributions are narrowed when the score is considered. This implies that the percentage of small factors ($| \alpha_{st} | < 4$) increases from 68.1% in MLC until 99.8% for S0p as can be seen in Table 2.

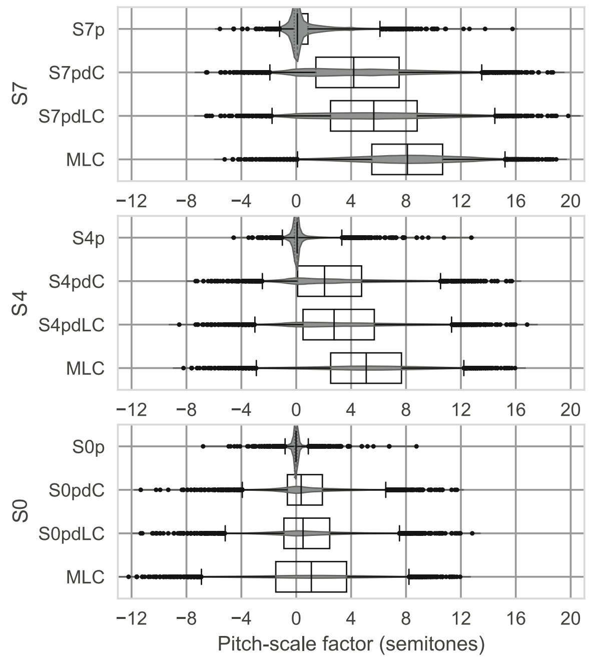


Fig. 5 Pitch-scale factors (α_{st}) for different vocal ranges (S0, S4, S7) and unit selection configurations with $T = 100$ bpm. Whiskers are set to 2nd and 98th percentile. Differences between all configurations are statistically significant ($p < 0.01$) except for the pair S0pdC-S0pdLC

Table 2 Pitch-scale factor ($| \alpha_{st} |$) percentages and good concatenation percentages

		$\| \alpha_{st} \|$				Concat.
	Configuration	[0–4]	(4–7]	(7–12]	> 12	Good
S7	S7p	94.2	4.5	1.2	0.1	33.1
	S7pdC (100 bpm)	48.0	24.0	22.9	5.1	67.5
	S7pdC (50 bpm)	47.1	24.0	23.7	5.1	68.1
	S7pdLC (100 bpm)	36.2	24.6	30.8	8.3	70.5
	S7pdLC (50 bpm)	36.2	24.1	31.2	8.4	70.4
	MLC	14.3	24.2	46.9	14.6	72.3
S4	S4p	98.6	1.2	0.3	0.0	44.2
	S4pdC (100 bpm)	69.2	19.0	11.1	0.7	70.4
	S4pdC (50 bpm)	68.7	18.9	11.7	0.7	71.2
	S4pdLC (100 bpm)	60.4	22.6	15.7	1.3	72.1*
	S4pdLC (50 bpm)	59.8	22.8	16.2	1.3	71.7
	MLC	37.7	31.4	28.6	2.3	72.3
S0	S0p	99.8	0.2	0.0	0.0	52.9
	S0pdC (100 bpm)	88.1	10.3	1.5	0.0	78.4
	S0pdC (50 bpm)	87.5	10.9	1.6	0.0	77.8
	S0pdLC (100 bpm)	82.5	14.2	3.3	0.0	76.7
	S0pdLC (50 bpm)	82.1	14.6	3.3	0.0	76.5
	MLC	68.1	25.6	6.3	0.0	72.3

Each row shows the percentages corresponding to a particular vocal range (S0, S4, or S7) and US configuration. Differences with respect to MLC are statistically significant ($p < 0.01$) for all configurations, except *

When singing beyond the speech corpus vocal range (S4 and S7), the distribution of MLC pitch-scale factors with S0 shifts up 4 semitones for S4 and 7 for S7 as seen in Fig. 5. Conversely, when the score is taken into account this increase can be mitigated, or even neutralised if only pitch is considered (Sxp). However, Table 2 shows that 72.3% of good concatenations obtained with MLC drop to 52.9% for S0p, 44.2% for S4p, and 33.1% for S7p. By contrast, the intermediate configurations (SxpdC and SxpdLC) still allow for a statistically significant reduction of the pitch-scale factors while minimising the concatenation quality degradation. Finally, it should be noted that in the score-driven configurations, the percentage of good concatenations decreases as the distance from the speech corpus vocal range midpoint increases.

5.1.2 Time-scale analysis

Regarding the time-scale factors, although the differences between configurations are in some cases statistically significant, they are barely relevant compared to the differences which arise from the tempo and the note values. According to this, and for the sake of clarity, the results of the intermediate configuration S4pdLC are presented for the two tempos under study, breaking them down according to the three most frequent note values: sixteenth note (𝅘𝅥𝅯), eighth note (𝅘𝅥𝅮), and quarter note (𝅘𝅥). These note values respectively account for 14.0%, 59.1%, and 21.7% of the notes in the score dataset, and they last 150 ms, 300 ms, and 600 ms for $T = 100$ bpm and double for $T = 50$ bpm.

Figure 6 shows the distributions of the time-scale factors (β), with the boxplot whiskers set to the minimum and the 98th percentile. The time-scale factor percentages by category are presented in the Table 3. We can see in Fig. 6 that when the tempo goes from 100 bpm to 50 bpm, notes doubled their duration, while time-scale factors more than doubled. This behaviour is also observed between note values within the same tempo. Similarly, Table 3 shows that while almost all (97.8%) of the shortest notes (150 ms) can be addressed with small time-scale factors ($\beta \leq 2.5$), when moving to medium duration notes (300 ms) 15.2% of high time-scale factors emerge at 100 bpm, and 17.4% at 50 bpm. Finally, as seen in Fig. 6 time-scale factors up to 28 can be required when singing long notes (600 ms), and even greater than 50 for notes lasting 1200 ms.

5.2 Subjective evaluation

Results from the MUSHRA test are shown in Tables 4 and 5. Regarding the singing assessment (see Table 4) the US-TTS&S framework has received MUSHRA scores of around 60. Although a slight preference for the contralto vocal range (S7) can be observed (62 at 100 bpm, and 61 at 50 bpm), similar results have been obtained for all the analysed scenarios.

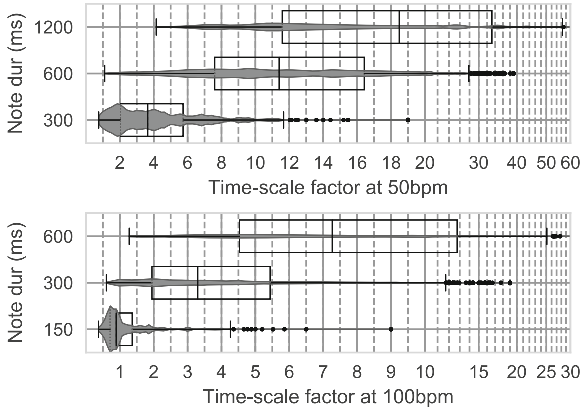


Fig. 6 Time-scale factors (β) obtained with the S4pdLC configuration at 100 bpm and 50 bpm for different note durations. Whiskers are set to minimum and 98th percentile

Table 3 Time-scale factor (β) percentages obtained with the S4pdLC configuration at 100 bpm and 50 bpm for different note durations (in ms)

Note	β			Note	β		
dur(ms)	≤ 2.5	(2.5–4]	> 4	dur(ms)	≤ 2.5	(2.5–4]	> 4
600	9.0	23.5	67.5	1200	0.0	0.2	99.8
300	55.1	29.7	15.2	600	3.5	8.4	88.1
150	97.8	1.8	0.3	300	50.1	32.6	17.4

With regard to naturalness (see Table 5), singing produced by the US-TTS&S framework is far from the Vocaloid (around 40 and 69, respectively). Although the differences between the US configurations are not statistically significant (according to the Wilcoxon signed-rank test), some tendencies can be observed. For instance, looking at the MUSHRA scores in Table 5 it can be seen that Sxp configurations have received the lowest ratings in all the analysed scenarios except for S4 and $T = 50$ bpm. Conversely, when the concatenation cost is enabled (SxpdC and SxpdLC), the naturalness is similar to that of MLC, or in some cases slightly improved, as with S0 and S0pdC at 50 bpm, or with S4 for both configurations and the two tempos.

5.3 Discussion

The experiments have been designed to evaluate the proposal through a proof-of-concept implementation. From the objective tests, it can be observed that large time transformation factors arise when dealing with medium duration notes (300 ms), but especially when long and very long notes (600 ms and 1200 ms) are present in the song (see Fig. 6). This result is in concordance with the corpus characteristics, which contains almost no vowels longer than 200 ms (see Fig. 4a). As a consequence, we can conclude that score-driven US configurations hardly impact on the time-scaling requirements.

Regarding pitch-scaling, the obtained moderate transformation factors required to sing in the speech corpus vocal range (S0) are consistent with the overlap between the $F0$ distribution from the score dataset and that from

Table 4 Singing MUSHRA average scores and 95% confidence interval

	Configuration	$T = 100$ bpm	$T = 50$ bpm
S7	MLC	*62 ± 6*	*61 ± 7*
S4	MLC	59 ± 6	60 ± 6
S0	MLC	60 ± 8	58 ± 7

Best values are in italics

Table 5 Naturalness MUSHRA average scores and 95% confidence interval

	Configuration	$T = 100$ bpm	$T = 50$ bpm
S7	V	74 ± 6	70 ± 6
	S7pdLC	41 ± 5	*44 ± 6*
	S7pdC	39 ± 6	43 ± 6
	S7p	36 ± 5	40 ± 6
	MLC	*42 ± 5*	*44 ± 5*
S4	V	69 ± 7	67 ± 7
	S4pdLC	*42 ± 6*	38 ± 6
	S4pdC	39 ± 6	*41 ± 6*
	S4p	35 ± 6	38 ± 6
	MLC	38 ± 6	37 ± 6
S0	V	66 ± 7	70 ± 6
	S0pdLC	*44 ± 5*	38 ± 4
	S0pdC	*44 ± 5*	*42 ± 6*
	S0p	41 ± 6	35 ± 6
	MLC	*44 ± 6*	39 ± 5

Best values achieved by the proposed system in each scenario are in italics

the corpus vowels (see Fig. 4). Conversely, when moving towards a contralto singer vocal range (S7), the overlap between $F0$ distributions is significantly reduced as it can be seen in Fig. 4. Thus, even though Sxp configurations are able to find almost all the vowels close to the desired pitch, it becomes harder to find units that also join adequately and meet the other target specifications (SxpdC and SxpdLC). This can be observed in the last column of Table 2, in the decreasing percentage of good concatenations when moving away from the corpus vocal range. Hence, although the score-driven US strategies have been proved helpful to reduce the pitch-scaling requirements, their effectiveness could be higher if a larger speech corpus with a greater coverage was available.

From the perceptual tests, a slight preference for singing in an actual singer vocal range (S7) has been observed (see Table 4). However, this preference is not significant with respect to the other vocal ranges under study (with MUSHRA scores of around 60). With regard to naturalness (see Table 5), the ratings achieved by the proof-of-concept with respect to natural speech are significantly different to those obtained by Vocaloid (with MUSHRA ratings around 40 and 69, respectively). Nevertheless, this is not surprising since Vocaloid is a commercial high-quality singing synthesiser exclusively designed for this purpose, which uses databases including diphones, sustained vowels and optionally triphones, sung by professional singers in several pitches to cover their vocal range [58]. Conversely, the proposal has to deal with the spoken units available in the corpus, which are low-pitched and very short compared to what could be found in a singing database. Therefore, converting them into singing involves high demanding transformations factors as seen in the objective evaluation. In this context, it has also been observed that the substantial pitch-scale factors reduction achieved by the score-driven US configurations has had a small impact on the naturalness, obtaining scores similar to those received by the text-driven US configuration. Besides the aforementioned restrictions due to the corpus size,this could be explained by the impossibility of relaxing the time-scale requirements. This may be important, considering that the ability to reproduce the behaviour of sustained vowels is known to be essential in singing synthesis [58].

Finally, it is worth mentioning that the validation of the proposal has been carried out with a specific speech corpus on a US-TTS system, since this approach enabled the study of the STS transformation factors required to produce singing from speech. Nevertheless, other corpus and adjustments of the cost function weights could be considered, and even other corpus-based approaches, such as statistical parametric speech synthesis using HMM or DNN.

6 Conclusions

This work has proposed a synthesis framework that provides singing capabilities to a US-TTS system from neutral speech, through the integration of speech-to-singing (STS) conversion. The proposal has been evaluated by means of a proof-of-concept implementation on a 2.6-h Spanish neutral speech corpus, considering different vocal ranges and tempos and studying diverse text-driven and score-driven US configurations.

Results show that high demanding STS transformation factors are required to sing beyond the corpus vocal range and/or when notes longer than 150 ms are present. However, the pitch-scale factors can be reduced by considering score-driven US configurations. Conversely, the time-scale requirements cannot be reduced because of the short length of the vowels available in the corpus.

Regarding the subjective evaluation, text-driven and score-driven US configurations have obtained a similar naturalness in all the analysed scenarios, with MUSHRA scores around 40. Although these values are far from those of Vocaloid (around 69), the obtained singing ratings around 60 validate the capability of the framework to address eventual singing needs.

The obtained results encourage us to continue working on the proposal to improve the performance of the system. To this aim, the focus will be placed on the generation of long sustained vowels, exploring advanced time-scale and spectral transformation strategies, and incorporating vibrato to the singing expression control generation module. Furthermore, other signal-processing techniques could be considered for the transformation module to better cope with the challenge of generating singing from neutral speech.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13636-019-0163-y.

Additional file 1: Singing S0 at 100 bpm with S0pdLC.
Additional file 2: Singing S4 at 100 bpm with S4pdLC.
Additional file 3: Singing S7 at 100 bpm with S7pdLC.
Additional file 4: Singing S0 at 50 bpm with S0pdLC.
Additional file 5: Singing S4 at 50 bpm with S4pdLC.
Additional file 6: Singing S7 at 50 bpm with S7pdLC.
Additional file 7: Singing S0 at 100 bpm with S0pdC.
Additional file 8: Singing S4 at 100 bpm with S4pdC.
Additional file 9: Singing S7 at 100 bpm with S7pdC.
Additional file 10: Singing S0 at 50 bpm with S0pdC.
Additional file 11: Singing S4 at 50 bpm with S4pdC.
Additional file 12: Singing S7 at 50 bpm with S7pdC.
Additional file 13: Singing S0 at 100 bpm with S0p.
Additional file 14: Singing S4 at 100 bpm with S4p.
Additional file 15: Singing S7 at 100 bpm with S7p.
Additional file 16: Singing S0 at 50 bpm with S0p.
Additional file 17: Singing S4 at 50 bpm with S4p.
Additional file 18: Singing S7 at 50 bpm with S7p.
Additional file 19: Singing S0 at 100 bpm with MLC.
Additional file 20: Singing S4 at 100 bpm with MLC.
Additional file 21: Singing S7 at 100 bpm with MLC.
Additional file 22: Singing S0 at 50 bpm with MLC.
Additional file 23: Singing S4 at 50 bpm with MLC.
Additional file 24: Singing S7 at 50 bpm with MLC.
Additional file 25: Synthetic spoken lyrics.
Additional file 26: Natural spoken lyrics.

Abbreviations

CBR: Case-based reasoning; DNN: Deep neural network; HMM: Hidden Markov model; MUSHRA: Multiple stimuli with hidden reference and anchor; NLP: Natural language processing; OLA: Overlap and add; STS: Speech-to-singing; TD-PSOLA: Time-domain pitch synchronous overlap and add; TTS: Text-to-speech; US-TTS&S: Unit selection-based text-to-speech-and-singing; US: Unit selection

Acknowledgements

We want to thank Yannis Stylianou for his advice regarding the HNM model, Jordi Bonada and Merlijn Blaauw for the Vocaloid samples, Lisa Kinnear for English proofreading, and the people that took the perceptual test.

Authors' contributions

All authors participated in the design of both the framework and its evaluation. MF and JCS carried out the proof-of-concept implementation, and MF conducted the experiments. The paper was mostly written by MF and FA. All authors have read and approved the final manuscript.

Funding

This work has been partially supported by the Agencia Estatal de Investigación (AEI) and FEDER, EU, through project GENIOVOX TEC2016-81107-P. Marc Freixes received the support of the European Social Fund (ESF) and the Catalan Government (SUR/DEC) with the FI grant No. 2016FI_B2 00094. Francesc Alías acknowledges the support from the Obra Social "La Caixa" under grants ref. 2018-URL-IR1rQ-021 and 2018-URL-IR2nQ-029.

Availability of data and materials

The data generated and analysed during the current study is available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Received: 26 November 2018 Accepted: 24 October 2019
Published online: 16 December 2019

References

1. P. Taylor, *Text-to-Speech Synthesis*. (Cambridge University Press, Cambridge, UK, 2009)
2. R. Montaño, F. Alías, The role of prosody and voice quality in indirect storytelling speech: Annotation methodology and expressive categories. Speech Commun. **85**, 8–18 (2016)
3. M. Fridin, Storytelling by a kindergarten social assistive robot: a tool for constructive learning in preschool education. Comput. Educ. **70**, 53–64 (2014)
4. J. Yamagishi, C. Veaux, S. King, S. Renals, Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. Acoustical Sci. Technol. **33**(1), 1–5 (2012)
5. L. Wood, K. Dautenhahn, B. Robins, A. Zaraki, in *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Developing child-robot interaction scenarios with a humanoid robot to assist children with autism in developing visual perspective taking skills, (2017), pp. 1055–1060
6. L. A. Jibb, K. A. Birnie, P. C. Nathan, T. N. Beran, V. Hum, J. C. Victor, J. N. Stinson, Using the MEDiPORT humanoid robot to reduce procedural pain and distress in children with cancer: A pilot randomized controlled trial. Pedia. Blood Cancer. **65**, 27242 (2018)
7. R. Khosla, K. Nguyen, M. T. Chu, Human Robot Engagement and Acceptability in Residential Aged Care. Int. J. Human-Comput. Interact. **33**, 510–522 (2017)
8. P. R. Cook, SPASM, a Real-Time Vocal Tract Physical Model Controller; and Singer, the Companion Software Synthesis System. Comput. Music J. **17**(1), 30–44 (1993)
9. Flinger. http://www.cslu.ogi.edu/tts/flinger/. Accessed June 2017
10. M. Uneson, Outlines of Burcas-A simple MIDI-to-singing voice synthesis system. Fonetik. **44**(1), 133–136 (2002)
11. M. W. Macon, L. Jensen-Link, J. Oliverio, M. A. Clements, E. B. George, in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*. A singing voice synthesis system based on sinusoidal modeling, vol. 1, (1997), pp. 435–438
12. J. Bonada, X. Serra, Synthesis of the singing voice by performance sampling and spectral models. IEEE Signal Process. Mag. **24**(2), 67–79 (2007)
13. M. Blaauw, J. Bonada, A neural parametric singing synthesizer modeling timbre and expression from natural songs. Appl. Sci. **7**(12), 1313 (2017). http://arxiv.org/abs/1704.03809
14. A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, WaveNet: A Generative Model for Raw Audio. arXiv 1609.03499, 1–15 (2016). http://arxiv.org/abs/1609.03499. Accessed Oct 2018
15. Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, R. A. Saurous, *Tacotron: Towards End-to-End Speech Synthesis* (ISCA, Stockholm, Sweden, 2017), pp. 4006–4010. https://doi.org/10.21437/Interspeech.2017-1452
16. S. King, Measuring a decade of progress in Text-to-Speech. Loquens. **1**(1), e006 (2014). https://doi.org/10.3989/loquens.2014.006
17. E. Del Blanco, I. Hernaez, E. Navas, X. Sarasola, D. Erro, in *Proc. Interspeech*. Bertsokantari: a TTS based singing synthesis system, (2016), pp. 1240–1244
18. M. Freixes, J. C. Socoró, F. Alías, in *Adv. Speech Lang. Technol. Iberian Lang. Third Int. Conf. IberSPEECH*. Adding singing capabilities to unit selection TTS through HNM-based conversion, vol. 10077 (LNAI, Lisbon, Portugal, 2016), pp. 33–43
19. A. Röbel, J. Fineberg, in *Proc. Interspeech*. Speech to chant transformation with the phase vocoder, (2007), pp. 2–3

20. T. Saitou, M. Goto, M. Unoki, M. Akagi, in *IEEE Workshop on Applicat. of Signal Process. to Audio and Acoust. (WASPAA)*. Speech-to-singing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices, (2007), pp. 215–218
21. M. Dong, S. W. Lee, H. Li, P. Chan, X. Peng, J. W. Ehnes, D. Huang, in *Proc. Interspeech*. I2R Speech2Singing perfects everyone's singing, (2014), pp. 2148–2149
22. J. Li, H. Yang, W. Zhang, L. Cai, A lyrics to singing voice synthesis system with variable timbre. Commun. Comput. Inf. Sci. **225**, 186–193 (2011)
23. J. M. Chowning, in *Sound Generation in Winds, Strings, Computers*. Computer synthesis of the singing voice (Royal Swedish Academy of Music, Stockholm, Sweden, 1980), pp. 4–13
24. P. R. Cook, Singing voice synthesis: History, current work, and future directions. Comput Music J. **20**(3), 38–46 (1996)
25. J. L. Kelly, C. C. Lochbaum, in *4th Int. Congr. Acoust*. Speech synthesis, (1962), p. 42
26. P. R. Cook, D. Kamarotos, T. Diamantopoulos, G. Philippis, in *International Computer Music Conference (ICMC)*. IGDIS (Instrument for Greek Diction and Singing): A Modern Greek Text to Speech/Singing Program for the SPASM/Singer Instrument, (1993), pp. 387–389
27. J. Sundberg, The KTH synthesis of singing. Adv. Cognit. Psychol. **2**(2), 131–143 (2006)
28. X. Rodet, Y. Potard, J.-b. Barriere, The CHANT Project: from the synthesis of the singing voice to synthesis in general. Comput. Music J. **8**(3), 15–31 (1984)
29. L. Feugère, C. D'Alessandro, B. Doval, O. Perrotin, Cantor Digitalis: chironomic parametric synthesis of singing. J. Audio EURASIP Speech Music Process. **2017**(2), 1–19 (2017). https://doi.org/10.1186/s13636-016-0098-5
30. The Festival speech synthesis system (2018). http://www.cstr.ed.ac.uk/projects/festival/. Accessed 2018-11
31. N. D'Alessandro, R. Sebbe, B. Bozkurt, T. Dutoit, in *13th Eur. Signal Process. Conf.* MaxMBROLA: A max/MSP MBROLA-based tool for real-time voice synthesis, (2005)
32. N. D'Alessandro, O. Babacan, B. Bozkurt, T. Dubuisson, A. Holzapfel, L. Kessous, A. Moinet, M. Vlieghe, RAMCESS 2.X framework—expressive voice analysis for realtime and accurate synthesis of singing. J. Multimodal User Int. **2**(2), 133–144 (2008)
33. K. Vijayan, H. Li, T. Toda, Speech-to-singing voice conversion: The challenges and strategies for improving vocal conversion processes. IEEE Signal Process. Mag. **36**(1), 95–102 (2019)
34. H. Kawahara, I. Masuda-Katsuse, A. De Cheveigné, Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Commun. **27**(3), 187–207 (1999)
35. A. Roebel, S. Huber, X. Rodet, G. Degottex, in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*. Analysis and modification of excitation source characteristics for singing voice synthesis, (2012), pp. 5381–5384
36. S. Huber, A. Roebel, in *Proc. Interspeech*. On glottal source shape parameter transformation using a novel deterministic and stochastic speech analysis and synthesis system, (2015), pp. 289–293
37. L. Cen, M. Dong, P. Chan, in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*. Template-based personalized singing voice synthesis, (2012), pp. 4509–4512
38. K. Vijayan, M. Dong, H. Li, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. A dual alignment scheme for improved speech-to-singing voice conversion (IEEE, Kuala Lumpur, 2017), pp. 1547–1555. https://doi.org/10.1109/APSIPA.2017.8282289
39. S. W. Lee, Z. Wu, M. Dong, X. Tian, H. Li, in *Proc. Interspeech*. A comparative study of spectral transformation techniques for singing voice synthesis, (2014), pp. 2499–2503
40. J. Sundberg, J. Bauer-Huppmann, When does a sung tone start? J. Voice. **21**(3), 285–293 (2007)
41. M. Umbert, J. Bonada, M. Goto, T. Nakano, J. Sundberg, Expression Control in Singing Voice Synthesis: Features, approaches, evaluation, and challenges. IEEE Signal Process. Mag. **32**(6), 55–73 (2015)
42. L. Formiga, A. Trilla, F. Alías, I. Iriondo, J. C. Socoró, in *Proceedings of FALA 2010, Jornadas en Tecnología del Habla and Iberian SLTech Workshop*. Adaptation of the URL-TTS system to the 2010 Albayzin evaluation campaign, vol. 1, (2010), pp. 363–370
43. A. Hunt, A. W. Black, in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP), vol. 1*. Unit selection in a concatenative speech synthesis system using a large speech database (IEEE, Atlanta, USA, 1996), pp. 373–376. https://doi.org/10.1109/ICASSP.1996.541110
44. F. Alías, L. Formiga, X. Llorà, Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept. Speech Commun. **53**(5), 786–800 (2011)
45. À. Calzada Defez, J. C. Socoró Carrié, Voice quality modification using a harmonics plus noise model. Cognit. Comput. **5**(4), 473–482 (2013). https://doi.org/10.1007/s12559-012-9193-9
46. D. Erro, A. Moreno, Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models. PhD thesis (2008)
47. Y. Stylianou, Applying the harmonic plus noise model in concatenative speech synthesis. IEEE Trans. Speech Audio Process. **9**(1), 21–29 (2001)
48. E. Nichols, D. Morris, S. Basu, C. Raphael, in *International Symposium on Music Information Retrieval*. Relationships Between Lyrics and Melody in Popular Music, (2009), pp. 471–476
49. D. Erro, A. Moreno, A. Bonafonte, in *6th ISCA Workshop on Speech Synthesis (SSW)*. Flexible harmonic/stochastic speech synthesis, (2007), pp. 194–199
50. F. Alías, X. Sevillano, J. C. Socoró, X. Gonzalvo, Towards high quality next-generation Text-to-Speech synthesis: a Multidomain approach by automatic domain classification. IEEE Trans. Audio, Speech Language Process. **16**(7), 1340–1354 (2008)
51. S. Karabetsos, P. Tsiakoulis, A. Chalamandaris, S. Raptis, One-class classification for spectral join cost calculation in unit selection speech synthesis. IEEE Signal Process. Lett. **17**(8), 746–749 (2010)
52. Cantos Infantiles Educativos, de Pablo Bensaya. http://presencias.net/cantos/kcanto.html. Accessed Oct 2018
53. H. Francois, O. Boeffard, in *Int. Conf. on Lang. Resources and Evaluation (LREC)*. The greedy algorithm and its application to the construction of a continuous speech database, vol. 5, (2002), pp. 1420–1426
54. E. Moulines, J. Laroche, Non-parametric techniques for pitch-scale and time-scale modification of speech. Speech Commun. **16**(2), 175–205 (1995)
55. G. P. Kafentzis, G. Degottex, O. Rosec, Y. Stylianou, in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*. Pitch Modifications of speech based on an Adaptive Harmonic Model, (2014)
56. R.ecommendation. ITU, ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems. Int. Telecommun. Union (2003)
57. N. Jillings, D. Moffat, B. De Man, J. D. Reiss, in *12th Sound and Music Computing Conference*. Web Audio Evaluation Tool: A Browser-based Listening Test Environment, (2015)
58. H. Kenmochi, in *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*. Singing synthesis as a new musical instrument, (2012), pp. 5385–5388

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.