Singing Voice Database

Liliya Tsirulnik[1](✉) and Shlomo Dubnov[2](✉)

[1] NTENT, San Diego, USA
liliya.tsirulnik@gmail.com
[2] University of California in San Diego, San Diego, USA
sdubnov@ucsd.edu

Abstract. The first publicly available singing voice database, which was first released in 2012, is presented in this paper. This database contains recordings of professional singers including one Grammy Award winner. The database includes so-called plain singing as well as singing with nine different singing expressions. For all the material there are both vocal and glottal voice recordings, where glottal recordings were made by placing the microphone on the neck of the singer near the glottis. Part of the database is annotated on the phoneme and pitch level, which makes it much easier to do automated analyses of different singing voice phenomena. Such varied content of the singing voice database makes it possible to use different types of singing voice for research, including interconnection of vocal and glottal singing voice signals, acoustic phenomena which take place in singing voice, different acoustic phenomena and effects of different expressions in singing voice, as well as comparing singing voice phenomena and acoustic effects of different singers. This database can also be used for simplified singing voice synthesis.

Keywords: Singing voice synthesis · Singing voice database · Speech Synthesis

1 Introduction

Singing Voice Synthesis Systems, as well as Speech Synthesis Systems, have been developed over several decades. Though the general approaches to Singing Voice Synthesis and Speech Synthesis are similar, singing voice is very different from spoken voice in terms of its production and perception by a human. Intelligibility of phonemic message in speech is very important, while in singing it is often secondary to the intonation and musical qualities of the voice. During singing voice synthesis it is important to convey singing voice phenomena, such as vibrato, jitter, drift, presence of singer's formant, and others.

Singing Voice Synthesis systems utilize different approaches, as presented in Sect. 2, using different amounts and different representations of voice units. But until recently, there was no publicly available singing voice database which would allow research on singing voice phenomena and formulate necessary and sufficient content of Singer's database for high quality singing voice synthesis. The present work describes the first (to the best of the authors' knowledge) annotated singing voice database, which was initially released in 2012. This database will allow the study of how various voice

A. A. Salah et al. (Eds.): SPECOM 2019, LNAI 11658, pp. 501–509, 2019.
https://doi.org/10.1007/978-3-030-26061-3_51

phenomena and effects are represented by spectral, temporal, and amplitude characteristics, as well as to create a simplified singing voice synthesis system.

The rest of the paper describes the database (Sects. 3 to 5), ending with our conclusions and possible uses of this database in Sect. 6.

2 Previous Work

Approaches to Singing Voice Synthesis systems include an articulatory approach [1], a formant synthesis [2, 3], and a concatenative and corpus-based synthesis [4–6]. In contrast to TTS-synthesis systems, where the input is a text and the output is a speech signal, for Singing Voice Synthesis systems the input is usually a musical score with lyrics, and the output is a synthesized singing voice signal.

The articulatory singing voice synthesis system SPASM [1] maps physical characteristics of the vocal tract to singing voice characteristics and produces a voice signal. The input to this system is not musical notes but the vocal tract characteristics. The system requires the user to have a knowledge of music and musical acoustics. For each note the user should specify seven parameters, including vocal tract shape (radius of each tract section), tract turbulence (noise spectrum and localization), performance features (random and periodic pitch), and others. The system takes into account singing voice phenomena, but the singing voice does not sound realistic.

The formant singing voice synthesizer CHANT [2] works with the English language. It is based on rules derived from signal and psychoacoustic analyses, such as the automatic determination of the formant relative amplitudes or bandwidths, or their evolutions depending on the variation of other external or internal parameters. CHANT uses an excitation resonance model to compose a singing voice signal. For each resonance, a basic response is generated using Formant Wave Functions, then these signals are summed to produce the resulting signal. The system's synthesis results are impressive in some cases, although it is said that this require tedious manual tuning of parameters.

Another formant singing voice synthesizer Virtual Singer [3] supports several languages, including French, English, Spanish, Italian, German, Japanese, and others. Virtual singer is an opera-like singing synthesizer. Its main attributes are the wide amount of languages that the synthesizer supports, the sound-shaping control (timbre and intonation), and the RealSinger function, which allows defining a Virtual Singer voice out of recordings of the user's own voice. The singer's database of Virtual Singer includes the set of phonemes with additional first parts of the diphthongs, represented as spectral envelopes. It assumes that only three to six formants are sufficient to generate a phoneme with acceptable quality. The advantage of this method is that only a small amount of data is required to generate a phoneme, and it is far easier to modify these data slightly to produce another voice timbre. However, the result is generally less realistic than with recorded speech elements.

The MaxMBROLA [4] is a concatenative synthesis system. It supports 30 languages and has 70 voice databases. MaxMBROLA is a real-time singing synthesizer based on the MBROLA speech synthesizer. It uses the standard MBROLA acoustic base which includes diphones and conveys singing voice phenomena by modifying the voice signal.

Another concatenative singing voice synthesis system – Flinger [5] – supports the English language. The singer's database of Flinger includes 500 segments of the consonant-vowel-consonant (CVC) structure: 250 on low pitch and 250 on high pitch, which is about 10 min of singing voice signal. The units are represented using Harmonic Plus Noise model. The system supports the following singing voice effects: vibrato, vocal effort, and variation of spectral tilt with loudness (crescendo of the voice is accompanied by a leveling of the usual downward tilt of the source spectrum).

Vocaloid [6] is currently considered as the best singing voice synthesizer for popular music. It supports the English and Japanese languages. It's a corpus-based system with possible pitch and duration changing with signal generation from the sinusoidal model. Singer's database includes natural speech segments. It should contain all the diphones (pairs of CV, VC, VV for English, where C is a consonant, V is a vowel) and can contain polyphones as well. The size of singer's base is 2000 units per one pitch.

3 Singing Voice Database Content

The main goal of creating the database was to represent different singing voice phenomena rather than a full set of phonemes for a particular language. That's why the a-priori database cannot be used for a full-fledged singing voice synthesis (where the input is a musical score and lyrics), rather for a simplified singing voice synthesis, where the input is a musical score.

The Singing Voice Database (SVDB) includes two parts: (1) Singing musical scale recordings and (2) Singing song recordings. The first part includes:

1.1. The scale (musical notes) performed using "ah" vowel ("ah-ah-ah" recordings)
1.2. The transitions between notes performed using "ah" vowel
1.3. The scale (musical notes) performed using "la" syllable ("la-la-la" recordings)
1.4. The transitions between notes performed using "la" syllable.

The second part includes just the song "Twinkle, twinkle, little star" [7].

Both parts contain plain recordings and recordings with special singing expressions described in Table 1. The database contains vocal recordings and also the so-called "glottal" recordings, which are made by placing the second microphone on a neck of the singer near the glottis.

4 Singing Voice Database Recording

All the recordings were performed in a studio by professional singers. For both the musical scale and the song one female and one male voice were recorded. For the musical scale the voices of Bonnie Lander [8] and Philip Larson [9] were recorded. For the song the voices of Grammy Award winner Susan Narucki [10] and Philip Larson were recorded. The singers' voice characteristics are given in Table 2.

Both vocal and glottal recordings were made simultaneously. The air microphone was used for vocal recordings and a contact microphone was used for glottal recordings.

The recordings are in WAVE PCM format with the following characteristics: 44100 Hz; 16 bit; 1 channel (mono).

Table 1. Singing expressions.

Expression #	Expression name	Description
1	Bounce	Increased articulation on consonants (slight increase of weight on initial consonants) followed by decrease of weight on adjacent vowels. More rhythmic vitality of a regular sort
2	Hollow	Less articulation of consonants. Modification of vowels to minimize their differences with the addition of "air" in the tone (as opposed to focused tone)
3	Light	Minimal initial articulation and weight. Modification of vowels to emphasize "brightness" upper partials
4	Soft	Modification of vowels; some air added, low volume. Consonants are present, but not sharply defined
5	Sweet	Extreme legato. Pure vowels. Consonants present but without extra articulated weight
6	Flat	Affectless. Consonants and vowels with same weight. Minimizing melodic contour
7	Mature	Emphasis on heavier vibrato in sound, (irregular) emphasis on lower partials of vowels (dark rather than bright)
8	Sharp	Emphasis on forward placement of vowel, cutting off lower partials. Aggressive articulation of consonants
9	Husky	Irregular rhythmic inflection in phrasing. Irregular pronunciation of consonants and vowels, additional throat grab noises and air to vowel mix
10	Clear	Purity of vowels and consonants. Emphasis on regularity of pronunciation. Sincere effect

Table 2. Singers voice characteristics.

Type of recordings	Singer	Gender	Voice	Musical notes range
Musical scale	Singer 1	F	Soprano	C4 to H5
Musical scale	Singer 2	M	Bass-baritone	C3 to H4
Song	Singer 3	F	Soprano	
Song	Singer 4	M	Bass-baritone	

5 Singing Voice Database Processing

The musical scale recordings were processed in the following way:

- pauses between recordings (unvoiced fragments) were automatically identified and marked,
- pitch annotation of voiced fragments was made.

For voiced/unvoiced fragments identification the following algorithm was used:

1. For each 5 ms audio frame with a step of 1 ms:

   1.1 The zero-crossing rate was calculated using the formula:

$$Z_n = \sum\nolimits_{m=0}^{N-1} \frac{|sgn[(x(n-m+1)]-sgn[x(n-m)]|}{2N} \tag{1}$$

where N – frame size,
$x(n)$ – signal at the n-th sample,

$$sgn[x(n)] = \left\{ \begin{array}{c} 1, x(n) \geq 0 \\ -1, x(n) < 0 \end{array} \right\}$$

1.2. The energy was calculated as a root-mean-square level:

$$E_n = \sqrt{\frac{1}{N}\sum\nolimits_{m=0}^{N-1} |x(n-m)|^2} \tag{2}$$

To calculate energy on each frame the Hamming window was used.

2. To smooth out the result the median value with the window size equal to 7 was calculated for each Energy and Zero-crossing rate.
3. The frame is considered to be voiced if

$$Z_n < Z_{th}\, and\, E_n \ > E_{th} \tag{3}$$

where Z_{th} is a zero-crossing threshold and E_{th} is an energy threshold. The values for threshold were chosen experimentally Z_{th} = 40 and E_{th} = 0.06

Pitch annotation algorithm was based on the fact that the recordings contain consequently sung notes. For example, the consequence of notes C4, D4, E4, F4 corresponds to the fundamental frequencies 261.63 Hz, 293.66 Hz, 329,63 Hz, 349,23 Hz. It means that the length of pitch period changes gradually. For automatic pitch annotation software the initial fundamental frequency (F_0) was specified manually. The software then finds and marks as a pitch period border the nearest zero crossing point in a singing voice signal, taking into account the length of the previous pitch period and voiced/unvoiced parts of the signal. The results were manually verified and corrected when needed.

For the second part of the recordings – song recordings – phoneme boundaries were semi-automatically found and all the vowel phonemes were annotated.

All the annotations were made in the TIMIT database files format [11]. For phonetic transcription the ARPABET code [12] was used. The phonetic transcription of the whole song as well as all the vowel phonemes marked in recordings are presented in Table 3.

Table 3. The lyrics and phonetic transcription of a song "Twinkle, twinkle, little star".

#	Word	Transcription	Vowel phonemes	#	Word	Transcription	Vowel phonemes
1	Twinkle	ˈT, W, IX, NG, K, L	IX	35	Blazing	ˈB, L, EY, Z, IX, NG	EY, IX
2	Twinkle	ˈT, W, IX, NG, K, L	IX	36	Sun	S, AH, N	AH
3	Little	ˈL, IX, T, L	IX	37	Is	IX, Z	IX
4	Star	S, T, AA	AA	38	Gone	G, AH, N	AH
5	How	H, AW	AW	39	When	W, EH, N	EH
6	I	AY	AY	40	There's	DH, AXR, Z	AXR
7	Wonder	ˈW, AH, N, D, AX	AH, AX	41	Nothing	ˈN, AH, TH, IX, NG	AH, IX
8	What	W, AH, T	AH	42	He	H, EE	EE
9	You	J, UX	UX	43	Shines	SH, AY, N, Z	AY
10	Are	AA	AA	44	Upon	AX,ˈP, AH, N	AX, AH
11	Up	AH,P	AH	45	Then	DH, EH, N	EH
12	Above	AX,ˈB, AH, V	AX, AH	46	You	J, UX	UX
13	The	DH, AX	AX	47	Show	SH, OW	OW
14	World	W, ER, L, D	ER	48	Your	J, AO	AO
15	So	S, OW	OW	49	Little	ˈL, IX, T, L	IX
16	High	H, AY	AY	50	Light	L, AY, T	AY
17	Like	L, AY, K	AY	51	Twinkle	ˈT, W, IX, NG, K, L	IX
18	A	AX	AX	52	Twinkle	ˈT, W, IX, NG, K, L	IX
19	Diamond	ˈD, AY, M, AX, N, D	AY, AX	53	Through	TH, R, UX	UX
20	In	IX, N	IX	54	The	DH, AX	AX
21	The	DH, AX	AX	55	Night	N, AY, T	AY
22	Sky	S, K, AY	AY	56	Twinkle	ˈT, W, IX, NG, K, L	IX
23	Twinkle	ˈT, W, IX, NG, K, L	IX	57	Twinkle	ˈT, W, IX, NG, K, L	IX

(*continued*)

Table 3. (*continued*)

#	Word	Transcription	Vowel phonemes	#	Word	Transcription	Vowel phonemes
24	Twinkle	ˈT, W, IX, NG, K, L	IX	58	Little	ˈL, IX, T, L	IX
25	Little	ˈL, IX, T, L	IX	59	Star	S, T, AA	AA
26	Star	S, T, AA	AA	60	How	H, AW	AW
27	How	H, AW	AW	61	I	AY	AY
28	I	AY	AY	62	Wonder	ˈW, AH, N, D, AX	AH, AX
29	Wonder	ˈW, AH, N, D, AX	AH, AX	63	What	W, AH, T	AH
30	What	W, AH, T	AH	64	You	J, UX	UX
31	You	J, UX	UX	65	Are	AA	AA
32	Are	AA	AA	66	In	IX, N	IX
33	When	W, EH, N	EH	67	The	DH, AX	AX
34	The	DH, AX	AX	68	Dark	D, AA, K	AA
69	Blue	B, L, UX	UX	85	The	DH, AX	AX
70	Sky	S, K, AY	AY	86	Morning	ˈM, AO, N, IX, NG	IX
71	So	S, OW	OW	87	Sun	S, AH, N	AH
72	Deep	D, EE, P	EE	88	Does	D, AH, Z	AH
73	Through	TH, R, UX	UX	89	Rise	R, AY, Z	AY
74	My	M, AY	AY	90	Twinkle	ˈT, W, IX, NG, K, L	IX
75	Curtains	ˈK, ER, T, N, Z	ER	91	Twinkle	ˈT, W, IX, NG, K, L	IX
76	Often	ˈAH, F, N	AH	92	Little	ˈL, IX, T, L	IX
77	Peep	P, EE, P	EE	93	Star	S, T, AA	AA
78	For	F, AO	AO	94	How	H, AW	AW
79	You	J, UX	UX	95	I	AY	AY
80	Never	ˈN, EH, V, AX	EH, AX	96	Wonder	ˈW, AH, N, D, AX	AH, AX
81	Close	K, L, OW, Z	OW	97	What	W, AH, T	AH
82	Your	J, AO	AO	98	You	J, UX	UX
83	Eyes	AY, Z	AY	99	Are	AA	AA
84	Till	T, IX, L	IX				

The resulting singing voice database has the following characteristics:

Part 1—“Ah-ah” and “La-la” recording of a male and a female voice. The overall length of the male voice recordings is 23 min and the female voice recordings is 33 min.

Part 2—song recordings of male and female voices. The overall length of the male voice recordings is 13 min and the female voice recordings is 15 min.

6 Conclusions

The singing voice database described here is publicly available from [13] and [14]. The database was first released in 2012 and is quite popular for research groups in Europe and America.

The advantage of the database created is that it includes not only so-called "plain" singing, but also singing with different expressions. It has both vocal and glottal recordings made simultaneously. It is partly annotated on pitch and phoneme levels. All these characteristics make it possible to use the database for different types of research, as well as for simplified singing voice synthesis.

Indeed, this database can be used to research different singing voice effects, including, but not limited to:

- interconnection of vocal and glottal singing voice signals,
- acoustic phenomena which take place in singing voice,
- different acoustic phenomena and effects of different expressions in singing voice, and
- comparison of singing voice phenomena and acoustic effects for different singers.

The first part of a database can be used for a singing voice synthesis as well. However, because it includes just "Ah-ah" and "La-la" sounds, it cannot be used for a full-fledged singing voice synthesis, as was mentioned before. But it can be successfully used for singing voice synthesis where the input is just musical notes (without lyrics).

References

1. Cook, P.R.: Singing Synthesis System. http://www.cs.princeton.edu/~prc/SingingSynth.html
2. Rodet, X., Potard, Y., Barrière, J.-B.: The CHANT project: from the synthesis of the singing voice to synthesis in general. Comput. Music J. **8**(3), 15–31 (1984)
3. Virtual Singer. http://www.myriad-online.com/en/products/virtualsinger.htm
4. MaxMBROLA. http://tcts.fpms.ac.be/synthesis/maxmbrola/description.php
5. Macon, M.W., Jensen-Link, L., Oliverio, J., Clements, M.A., George, E.B.: A singing voice synthesis system based on sinusoidal modeling. In: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, pp. 348–352. IEEE Computer Society Press (1997)
6. Kenmochi, H., Ohshima, H.: Vocaloid - commercial singing synthesizer based on sample concatenation. In: 8th Annual Conference of the International Speech Communication Association, ISCA, Antwerp, Belgium, pp. 87–88 (2007)
7. "Twinkle, twinkle, little star" song. https://en.wikipedia.org/wiki/Twinkle,_Twinkle,_Little_Star
8. Lander, B.: http://www.bonnielander.com/p/about.html

9. Larson, P.: https://music-cms.ucsd.edu/people/faculty/regular_faculty/philip-larson/index.html
10. Narucki, S.: http://www.susannarucki.net/home
11. TIMIT. https://catalog.ldc.upenn.edu/LDC93S1
12. ARPABET. https://en.wikipedia.org/wiki/ARPABET
13. Singing Voice Database. https://liliyatsirulnik.wixsite.com/svdb
14. Singing Voice Database. http://crel.calit2.net/projects/databases/svdb