Proposal for Arpasing, a Naming Standard and English Recording Scripts for UTAU the Singing Synthesis Software

Kanru Hua

October 5, 2016

1 Outline

This document proposes Arpasing, a *naming standard* based on Arpabet phoneme set and a set of phonetically balanced *recording scripts* for English speech databases (as known as “voicebank”), for use in UTAU, a singing synthesis software.

It should be pointed out in the first place, that the naming standard and recording scripts presented in this document share the same name, but are separately maintained. The recording scripts here presented are designed for *an implementation, in the form of a speech database, complying with the Arpasing naming standard*, and are not the unique solution for such purpose.

2 Naming Standard (version 0.1.1)

As suggested by its name, Arpasing adopts a slightly modified version of Arpabet phoneme set[1]. Such choice of phoneme set is justified by the availability of resources (e.g. pronunciation dictionary and corpora, mainly contributed by Carnegie Mellon University) and its ease to learn, as opposed to IPA or XSAMPA.

2.1 File Naming

This section specifies the naming of sound files in the database directory. Names of sound files under the same directory can take either of the two forms: direct naming or indexed naming.

2.1.1 Direct Naming

Direct naming simply associates the sound file with its phoneme transcription in lower case, delimited by underline “_”. This is essentially the same as what’s been applied in most non-Japanese UTAU speech databases, where sequences of phonemes or syllables of the target language are used to name the sound files.

Note that Arpasing does **not** allow heading or trailing underline before the file extension (e.g. `_x_x_x_.wav`).

2.1.2 Indexed Naming

Indexed naming is for the cases where the phoneme transcription is too long that might exceed the file system’s restriction on file name length, or simply too long to be human-readable in a typical file manager. Key-value pairs of file names and corresponding lower-cased phoneme transcription are stored in `csv` format in `index.csv` file under the directory containing sound files. The phoneme transcription is delimited by underline, which is consistent with the direct naming case (except for the absence of file extension). A typical entry in `index.csv` may look like

```
0.wav,ao_th_er_ah_v_dh_ah_d_ey_n_jh_er_t_r_ey_l
```

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Naming of the sound file can be arbitrary as long as there exists a corresponding entry in `index.csv`. Even though it is *recommended* to name the sound files in ascending numerical order (in accordance with the recording script) starting from zero.

2.2 Unit Naming

We define units as continuous portions of sounds in the speech database to be concatenated by the speech synthesizer, each associated with a **unique** name relevant to its phonetic representation. As for UTAU, units are defined in `oto.ini`, an index file distributed with speech databases and the name of an unit is called alias.

Arpasing has a loose restriction on the number of phonemes in a unit. Units can be diphones, triphones, quadphones or even a combination of all. However for triphones and longer units, only certain forms of phonetic structure are allowed.

2.2.1 Diphones

Diphones are the most basic form of units consisting of two lower case phonemes delimited by a space character, regardless of the phonetic class (e.g. consonant or vowel) of either phoneme. Heading or trailing silence is represented by dash ("-"). The following lists examples of legitimate diphone unit names
`- ao`, `er ah`, `v dh`, `dh ah`, `ey n`, `l -`
Note that diphone units are still required no matter if the database contains triphone or longer units.

2.2.2 Triphones (optional)

Triphone units take the form of "any-consonant-any". The first and last phoneme can be either consonantal or vowel. The following lists examples of legitimate triphone unit names
`ao th er`, `ah v dh`, `v dh ah`

2.2.3 Quadphones (optional)

Quadphone units are the direct extension of triphone units, taking the form of "any-consonant-consonant-any". The following lists examples of legitimate quadphone unit names
`ih k sh ah`, `ow s t er`, `ih k t s`, `n t s -`

2.2.4 Suffix

Arpasing supports the use of upper case note name (in chromatic scale) as suffix of unit name. `C#`, `D#`, `F#`, `G#`, `A#` shall be used in place of `Db`, `Eb`, `Gb`, `Ab`, `Bb`. This is also consistent with most existing UTAU databases. Prefixes are not supported.

2.2.5 Handling of duplicated units

Duplicated occurrences of the same unit are suggested **not** to be removed as they offer users alternatives which may fit better in different musical and linguistic contexts, and ultimately, facilitate a somewhat laborious form of unit-selection. To prevent naming conflict, starting from the second occurrence an 1-based counting integer is inserted between unit name and suffix (if exists). The order is not important, but should be *consistent across samples recorded at different pitches* in the case of multi-pitch voicebanks. Number shall **not** be attached to the first occurrence of a unit.

Note: to ensure ordering consistency for multi-pitch voicebanks, indexed naming (2.1.2) is highly recommended over direct naming (2.1.1).

3 Recording Script (version 0.1)

3.1 Design Goal

In its simplest form, the design goal of a speech synthesis corpus is to *cover most phonetic context with as little recording effort as possible.* Typical speech synthesizers may also require nice coverage on triphones, syllables (depending on language), and even semantic contexts to take account of the intonation variation.

Nicely, the case can be simplified for singing synthesis since music is encoded in a much more structured manner than speech in general, and acoustic features of the voice is less affected by semantics, while the pitch becomes highly dependent on the score. A fancy way to summarize this is that timbre, pitch and duration become more orthogonal. In our case the availability of a good-enough speech modification algorithm lessens the need for pitch and duration variety; our focus shifts to the major factors affecting timbre, that is to say, the local phonetic context.

The following recording script consists of two parts: the first part being designed for the most frequent diphone (or bi-gram) coverage while the second is an optional extension for better n-gram ($n \geq 3$) coverage, based on statistics collected from lyrics of 2000 English songs selected from the top 100 list from 1990 to 2011. Each line (except for the first 10 lines for vowels) contains three syllables with the same vowel. Lines with the same vowel are grouped together for better consistency.

3.2 Part 1: Basic Monophone/Diphone Coverage

This part of the recording script achieves 96% diphone frequency coverage[2], even though it only covers less than half of all possible diphones in English.

The way to read this list. For each line in the "phonetic transcription" column read all syllables at once without pausing. Pronunciation of the words in the "pronunciation guide" column is similar to that of the second column, but exact equivalence is not always guaranteed[3] and hence the pronunciation guide should only be used as a hint just to get a quick idea of how the syllables sound like.

No.	phonetic transcription	pronunciation guide (reference only)
0	ay - er - iy - ah - iy	
1	ih - iy - ay - ih - uw	
2	ow - ih - uw - ah - ow	
3	ah - iy - ih - ow - ay	
4	uw - ih - uw - ay - ey	
5	ey - ih - er - aw - er	
6	iy - ey - ay - ow - ah	
7	ao - aw - ay - ay - ae	
8	eh - er - ah - uw - aa	
9	er - ay - ah - eh - aw	
10	p aa r k - d aa r - y aa l	parc-dar-ya'll
11	k aa r v - w aa ch - l aa r jh	carve-watch-large
12	z aa n - ch aa r m - f aa r	zan-charm-far
13	g aa t - hh aa r t - t aa n	got-hardt-tonn
14	aa r m d - jh aa k - hh aa s	armed-jock-haas
15	m aa b - s aa l v - s aa f t	mob-solve-soft
16	l aa r m - w aa d - d r aa p	laarm-wad-drop
17	n aa k - q aa z - b aa m	nak-'az-balm
18	s k ae n - dh ae t - m ae n s	scan-that-mance

[2] We define the term "diphone frequency coverage" as certain percentage of all diphones (including duplicates) in the source corpus being covered. Here the 96% diphone frequency coverage means that 96% of the diphones in the 2000 song lyrics corpus can be found in part 1 of the recording script.

[3] In fact the pronunciation guide is generated by a program.

19	hh ae v - b ae k t - ae d z	halve-backed-ad’s
20	l ae f s - g r ae s - y ae m	laughs-gras-yam
21	jh ae ng - f ae k t - th ae ng k s	jang-fact-thank’s
22	s ae ng - y ae m k - q ae sh t	sang-yamk-’ashed
23	d ae l - y ae ng - s k ae l p	dal-yang-scalp
24	g l ae n - n ae p - ae z	glahn-nap-as
25	k ah m z - p ah - r ah sh	comes-pah-rusch
26	t ah ng - ah - hh ah n t	tongue-AH-hundt
27	p r ah - w ah s - m ah ng	prah-wass-mahng
28	ch ah ng - b ah - th ah n	chung-bah-thun
29	g ah - dh ah - jh ah n t	gah-the-jundt
30	z ah m - n y ah - d ah ch	zahm-nyah-duch
31	dh ah s - ah t - n ah l z	thus-utt-nahlz
32	k ah p s - f l ah - b ah v	cupps-flah-bahv
33	sh ah f - f ah k - sh ah z	schuff-fuck-shahz
34	v ah l dh - q ah m - hh ah g z	vahld-’umm-hugs
35	s p ao r t s - hh ao l - r ao ng d	sport’s-hall-wronged
36	s k ao r - l ao - f ao r b z	scor-law-forbes
37	t ao k t - f ao r d - y ao l	talked-foard-y’all
38	q ao r p s - n ao r th - d ao r d	’orps-north-daord
39	th ao t - s ao ng z - g l ao s t	thought-song’s-glossed
40	w ao r m th - b ao l - g ao r	warmth-ball-goar
41	r ao ng - m ao n - w ao r n d	rong-maune-warned
42	b aw z - m aw dh z - dh aw	boughs-mouths-thao
43	t aw n - p r aw l - l aw d	town-prowl-loud
44	f aw l k - q aw z - th aw t	foulk-’auz-thuot
45	n aw n - hh aw s - d aw t	noun-house-doubt
46	z ay - g ay - hh ay n d	zay-gae-hind
47	q ay m - m ay t - r ay k s	’ime-might-reich’s
48	f ay s - s l ay - w ay f	feis-sligh-wife
49	d ay d - s t ay l - n ay n th	died-stile-ninth
50	k r ay b - s k ay - r ay p s	krayb-sky-rayps
51	v ay v d - s ay n - b ay z	vayvd-sein-bies
52	p ay - jh ay - sh ay n	pie-jai-shine
53	n eh n - m eh k - g eh	nehn-mech-geh
54	f l eh sh - v eh l - k eh r	flesh-vehl-care
55	hh eh l - w eh l m d - q eh g z	hell-wehlmd-’eggs
56	y eh s - s t eh d - b eh l t	yes-stead-belt
57	p eh r z - sh eh - t w eh n	pairs-sheh-twehn
58	s eh p t - d r eh m t - d eh v	sept-dreamt-dev
59	sh eh f s - dh eh r - f eh	chef’s-their-feh
60	p w eh r - r eh k t - ch eh r z	pwehr-recht-chairs
61	er v - dh er - w er th	irv-thur-werth
62	m er - dh er d - sh er	murr-dherd-scher
63	z er - hh er t s - ch er p	zer-hirtz-chirp
64	s er d - w er l d z - v er s	serd-world’s-vers
65	g er z t - q er n d - hh er t	gerzt-’earned-herdt
66	n er v - t er - y er z	nerve-ter-yerz
67	s l er p - y er - f er g	slurp-yer-ferg
68	d er z - f er m - dh er z	derz-ferm-dherz
69	p er k - b er n - k er b	perc-bern-curb

70	hh ey l - s r ey - sh ey p	hail-srey-shape
71	ch ey n jh d - p ey v d - f ey d z	changed-paved-fades
72	s ey f - w ey z - ey m	safe-wais-aim
73	ey - dh ey - y ey	EY-they-yay
74	n ey k s - k ey v d - g ey jh	neyks-caved-gage
75	p l ey n z - k r ey - hh ey v	plaines-cray-hheyv
76	f ey th - b ey k - q ey v	faith-bake-'ave
77	m ey - t ey s - d ey b	mae-teys-deyb
78	v ey g - f ey - w ey v	vague-fay-wave
79	s w ih f t - k ih l - q ih g z	swift-kill-'iggs
80	z ih p - m ih - v ih n s	zip-mih-vince
81	n ih sh t - ng ih ng - w ih th	nihsht-ngihng-withe
82	p ih d - n ih k - w ih dh	pihd-knick-with
83	d ih k t - dh ih s - b r ih m	dihkt-this-brim
84	th ih n - f ih r - s ih z	thin-fear-sihz
85	s t ih - t l ih ng - l ih f t	stih-tlihng-lift
86	b r ih ng - dh ih n - sh ih f t	bring-dhihn-shift
87	ch ih ng - g ih r - dh ih	ching-gear-dhih
88	y ih r - hh ih ch t - jh ih g	year-hitched-jig
89	p ih m p - b ih n - r ih jh	pimp-been-ridge
90	p iy - w iy d - k iy m	p-we'd-keim
91	iy z - r iy m - l iy sh	e's-ream-leash
92	p l iy z d - m iy t - b iy d z	pleased-meat-beads
93	v iy - jh iy - ch iy v	v-g-chiyv
94	sh iy - b r iy dh d - dh iy	she-breathed-thee
95	iy - y iy - g iy	IY-ye-ghee
96	s iy k - p l iy - f iy s t	seek-plea-feast
97	z iy - hh iy t - n iy th	xie-heat-niyth
98	hh iy r z - s t iy n k - q iy v z	hears-steenk-'eves
99	b l iy p - d iy l z - w iy l	bleep-deal's-we'll
100	g r ow v - m ow k - m ow s	grove-moacq-mows
101	k ow p - dh ow - b ow	cope-tho-beau
102	k l ow dh z - hh ow m d - q ow d	clothes-holmd-'owed
103	t ow t - p ow - w ow k	tote-po-woke
104	s ow - y ow - sh ow l	so-yau-schaul
105	ng ow - hh ow - f ow n	ngow-hoe-phone
106	d ow n t - g ow z - n ow z	don't-goes-knows
107	b oy z - jh oy n - g oy	boies-join-goy
108	p uh t - sh uh r - w uh l f	put-schuur-wolf
109	g uh d z - y uh ng - t uh r z	good's-jung-tour's
110	b uh sh - k uh d - l uh k	busch-could-look
111	f y uw - d uw m - b uw s t	few-doom-boost
112	v y uw z - l uw k - k y uw b	views-luke-cube
113	m uw dh - m y uw - sh uw t	muwdh-mew-shoot
114	p uw r - g r uw v - r uw f	poor-groove-roof
115	th r uw - g r uw - y uw th	threw-grew-youth
116	f r uw t - q uw l d - t uw n	fruit-'ooled-toon
117	b y uw - w uw p - f uw l d	byuw-whoop-fooled
118	s uw dh - y uw z - n uw	soothe-ewes-gnu
119	k uw - hh uw m - g r uw p	coo-whom-group

3.3 Part 2: N-gram Coverage (Optional)

The following list raises triphone frequency coverage from 27% to 42%.

No.	phonetic transcription	pronunciation guide (reference only)
120	k aa s t - w aa n t - s t aa r t	cost-want-start
121	p r aa m - aa n - m aa m z	prom-on-mom's
122	w aa z - s t aa p - k aa r d z	waas-stop-card's
123	s aa r - b aa - d aa n t	saar-bah-daant
124	n aa t - g aa n - s k aa r d	knot-gohn-scarred
125	k ae n t - hh ae n d z - dh ae t s	can't-hand's-that's
126	y ae - y ae n - y ae	yeah-yahn-yeah
127	f ae s t - k ae n t - s ae	fast-can't-sae
128	dh ae t - y ae m - m ae t	that-yam-mat
129	z ae k - dh ae t - b ae ng k	zach-that-banc
130	b l ae s t - s t ae n d z - b r ae n d	blast-stands-brand
131	hh ae d - dh ae t - dh ae n	had-that-than
132	s l ah m z - dh ah - l ah s t	slums-the-lust
133	k r ah m - dh ah s - hh ah m p	crum-thus-hump
134	b ah m - sh ah n d - ah n t	bum-shunned-ahnt
135	z ah n d - dh ah - f r ah n	zahnd-the-frahn
136	t ah ch t - s ah k t - ah p	touched-sucked-up
137	b l ah n - dh ah - m ah n z	blahn-the-munns
138	m ah n - l ah v - m ah d	mun-love-mud
139	s ah n t - m ah n d - t ah l d	sundt-mund-tahld
140	d ah n d - n ah s t - w ah t	dunned-nahst-what
141	b ah t s - dh ah m - k l ah b	but's-dhahm-club
142	m ah n - ah n d - k ah n d z	mun-and-kahndz
143	m ah s t - f r ah n t - ah s t	must-front-ahst
144	f r ah m - w ah n s - b ah t s	from-once-but's
145	t r ah s t - ah v - dh ah m	trust-of-dhahm
146	p ah n d - y ah n - w ah l	pahnd-youn-wahl
147	z ah n t s - jh ah s t - s l ah g	zahnts-just-slug
148	s t ah d z - g l ah n d - w ah t s	studds-glahnd-what's
149	s ah n d - s ah m - f l ah d z	sund-some-floods
150	w ah t - y ah ng - d r ah ng k	what-young-drunk
151	jh ah s t - ah n d - l ah v d	just-and-loved
152	ah n t - dh ah - w ah t	ahnt-the-what
153	m ao l - y ao r - s t ao l d	mall-yore-stalled
154	m ao r n - y ao r k - f l ao r	morn-york-floor
155	sh ao r t - s ao l t - l ao ng	short-salt-long
156	p ao r - y ao r - f ao r m	por-yore-form
157	hh ao n t - y ao r - hh ao r n z	haunt-yore-horn's
158	f ao l z - ao l - w ao k t	fall's-all-walked
159	g r aw n d - b aw t - b aw n s	ground-'bout-bounce
160	k aw n t - n aw n - d aw n	count-noun-down
161	ay m - l ay z d - f r ay d	i'm-layzd-fried
162	th r ay v - f ay n d - m ay s	thrive-find-meiss
163	m ay l d - m ay n d - ay m	mild-mind-i'm
164	w ay l - l ay t s - n ay t	weil-light's-knight
165	d ay s - m ay - l ay f	deiss-mai-life
166	r ay - t ay m z - l ay k t	rye-time's-leicht

167	k r ay s t - l ay n - m ay l d	christ-line-mild
168	k ay t - m ay - m ay n d	kight-mai-mind
169	ay l - m ay n - s ay d z	aisle-mine-side’s
170	b eh t s - g eh n - s t eh r z	bet’s-’gain-stairs
171	f r eh n d z - g eh t s - s eh l f	friend’s-gets-self
172	d eh n t - l eh t - m eh n d	dent-let-mend
173	dh eh m - b eh g d - s eh k s	them-begged-sex
174	l eh f t s - t eh n t - t r eh s	left’s-tent-tress
175	s w eh p t - s p eh n d - dh eh n	swept-spend-then
176	s eh n - eh v - r eh s	sen-ev-ress
177	w eh s t - dh eh r z - b eh s t	west-theirs-best
178	d r eh s t - t eh l - s eh d	dressed-tel-said
179	m eh n d - w eh n - y eh l	mend-wen-yell
180	y er n - d er s t - hh er d	yearn-durst-heard
181	f l er t - f er s t - g er l	flirt-first-girl
182	b ey - b ey t - m ey k s	bay-bait-makes
183	b r ey n z - p l ey s t - b r ey k	brain’s-placed-brake
184	s t ey n d - t ey k s - t r ey n t	stained-takes-treynt
185	t ih n - dh ih - s ih z	tin-dhih-sihz
186	l ih t s - t r ih l - s k ih n	lits-trill-skin
187	d r ih ng - m ih n - l ih s	dring-mihn-lis
188	k ih ng - ih n - t ih ng z	king-in-tihngz
189	s l ih p s - d ih ng - t ih t	slips-ding-tit
190	s ih n - p r ih t - k r ih p s	sin-pritt-cripps
191	ng ih ng - hh ih t - b ih ch	ngihng-hit-bitch
192	v ih ng - y ih r z - s ih t	vihng-year’s-sit
193	s p ih n - s ih m - th ih ng z	spin-sim-thing’s
194	n ih t - w ih t s - w ih sh t	knit-wit’s-wished
195	g ih v - z ih k s - p ih t	give-zihks-pit
196	w ih - l ih n - m ih s t	wih-lin-missed
197	s t ih l - w ih n d - ih t s	stihl-wihnd-it’s
198	th ih ng k - ih t s - ih ng	think-it’s-ing
199	p ih ng - n ih ng - s t ih r	ping-ning-stear
200	ih t s - b ih l d - b ih k	it’s-bild-bic
201	f r iy k t - w iy r - m iy t s	freaked-we’re-meats
202	f iy l d - b iy s t - m iy n	feild-beast-mean
203	k iy p - d r iy m - n iy d z	keep-dream-needs
204	s l iy v - m iy l - m iy z	sleeve-meal-mease
205	w iy k s - k r iy m z - s w iy t	weaks-creams-suite
206	k l iy n - t r iy t s - m iy	clean-treats-me
207	f l ow n - l ow n - t r ow l d	flown-loan-trowld
208	w ow n t - w ow n t - s ow l d	won’t-won’t-sold
209	hh ow s t - d ow n t - b ow t	host-don’t-boat
210	d ow n t - n ow n - hh ow l	don’t-known-hoel
211	y uh r - t w uh d - y uh r z	you’re-twuhd-yours
212	y uw z d - t uw m - s t uw p	used-tomb-stoop
213	t uw d - k y uw - g r uw m	tude-cue-groom
214	f uw l - y uw - n uw z	fool-ewe-news
215	n uw n - y uw m - p y uw	noon-yuwm-peugh
216	p y uw k - d y uw d z - y uw s	puke-dudes-use
217	l uw n - t uw - s t uw	loon-tew-stew

218	y uw - t uw t - t uw l z	ewe-toot-tools
219	m uw v - g uw n z - t uw b	move-goons-tube

4 Roadmap

Implementation of Arpasing would be a joint effort by UTAU users and researchers/developers. Tasks to facilitate efficient creation and accessing of Arpasing-conforming databases include,

- Build the first Arpasing-conforming speech database
- Build a reference database using the proposed recording script
- Create tools for automatic labelling of voice samples
- Create tools for database querying and grapheme-to-phoneme conversion
- Create tools for database creation and management
- Write tutorials on building and using an Arpasing-conforming speech database in UTAU
- Extend this standard and recording script to cover accents other than North American English

Some of the tasks listed above are inter-dependent. For example, creating the first Arpasing database would be much simpler if an automated labelling tool is available; while the testing of tools in turn depends on at least one existing Arpasing database. The author suggests the following workflow as the plan for the next 6 months,

1. UTAU user community provides the first recording of both parts (3.1 and 3.2) of the proposed script.
2. The author gives feedback on pronunciation; the recording as well as this document will be revised.
3. Tool for phrase-level speech segmentation will be developed; diphone Arpasing will be supported by Moresampler 0.8.0.
4. Tools for database querying and editing will be developed.
5. UTAU user community releases the first Arpasing database as the reference database.
6. Tutorials relevant to Arpasing will be written and released.

Change log

of the naming standard

0.1.1 Fix typo in section “Indexed Naming” where phonemes shouldn’t be upper-cased.
0.1.0 First version of Arpasing naming standard.

of the recording script

0.1.0 First version of Arpasing recording script.

References

This work was inspired by the following publications,

[1] J. Matoušek, P. Josef and K. Jiří, “Design of speech corpus for text-to-speech synthesis”. in *Eurospeech 2001, Scandinavia.*

[2] J. Bonada, “Voice processing and synthesis by performance sampling and spectral models”. PhD Thesis. Universitat Pompeu Fabra, 2008.

[3] J. Kominek and A. W. Black, “The CMU Arctic speech database”. in *Fifth ISCA Workshop on Speech Synthesis*. 2004.