CHATR the Corpus; a 20-year-old archive of Concatenative Speech Synthesis

Nick Campbell

Speech Communication Lab,
School of Computer Science & Statistics
Trinity College Dublin, Ireland
nick@tcd.ie

Abstract

This paper reports the preservation of an old speech synthesis website as a corpus. CHATR was a revolutionary technique developed in the mid nineties for concatenative speech synthesis. The method has since become the standard for high quality speech output by computer although much of the current research is devoted to parametric or hybrid methods that employ smaller amounts of data and can be more easily tunable to individual voices. The system was first reported in 1994 and the website was functional in 1996. The ATR labs where this system was invented no longer exist, but the website has been preserved as a corpus containing 1537 samples of synthesised speech from that period (118 MB in aiff format) in 211 pages under various finely interrelated themes The corpus can be accessed from www.speech-data.jp as well as www.tcd-fastnet.com, where the original code and samples are now being maintained.

Keywords: ATR, CHATR, speech synthesis, corpus, preservation, concatenative synthesis, waveform samples, aiff, multilingual

1. Introduction

The CHATR speech synthesis system was developed throughout the early nineties in Kyoto, Japan, by researchers in Department 2 of the now defunct ATR Interpreting Telephony Labs (later Interpreting Telecommunications Research Labs) and was announced in 1996 as "a high-definition speech re-sequencing system" at the joint ASJ/ASA meeting in Hawaii [1], though the basic method was first reported in 1994 at the ESCA/IEEE Mohonk Speech Synthesis workshop [2]. The name was derived from "Collected Hacks from ATR" and was first suggested by Paul Taylor who was then working on the intonation component. It was not the first concatenative speech synthesis system but it was the first to use raw waveform segments directly, without recourse to any signal processing. This step not only greatly simplified the synthesis process but also allowed the use of very high quality recordings (some even in stereo) that exactly reproduced the voice quality and speaking style of the recorded subjects. It replaced the buzzy artificial sound of parametric synthesis with surprisingly natural-sounding speech. It was susceptible to concatenation errors if the waveform coverage in the voice database was incomplete but in that period much progress was made using as little as one hour of recorded speech and the samples in the corpus are all produced from such small databases. In contrast, some commercial users of this system now employ corpora of well-over 100 hours of recordings.

2. The CHATR Corpus

The CHATR Corpus has ISLRN 074-692-309-096-0 (registered on Oct 16 2015) and is listed under the URL: http://netsoc.fastnet.ie/chatr/. There are copies under http://www.tcd-fastnet.com and http://www.speech-data.jp. The corpus is freely available as an Open Access resource (Gratis & Libre) for research use and as a historical archive under CC-BY (Attribution) licensing. Figure 1 shows the top page and CHATR logo.

3. Page Layouts

This archive maintains the original linking structure (which has been superceded by more transparent structures in recent-day web-based resources), but this paper is intended to serve as a guide to what the corpus contains.

As shown in Figure 2, each page is provided with a set of arrows at the bottom for navigation and contains links to related sub-pages or sound samples. The arrows provide 'forward' and 'back' links with a central link for returning to a 'home' or a parent page. The arrows provide a slide presentation type of tour through the pages but there was no clear structure or overall view that allowed easy jumping between the various sections. These were the early days of the internet and the pages were produced more as a resource for presenting to visitors than as a stand-alone web archive for unsupervised remote access.

The pages were originally produced in Japan for mixed nationality and mixed-background audiences. They include many sections using Japanese characters and fonts but the headers were designed for an international audience. There is a whole subsection of the corpus under the "e_tour" subdirectory that was designed for presentation to English-speaking audiences with a more technical interest. The audio files include samples in four languages.

Figure 1: the archive's top page

The main sections were designed to introduce non-specialists to the technique of speech synthesis and to show the improvements offered by the concatenative raw-waveform variety. They include a review of then current synthesis samples, a description of the concatenative synthesis process, examples of contrastive prosody giving different meanings to the same text input, and many samples of different voices, young and old, famous and unknown, speaking a variety of languages that their original owners were probably not even familiar with.

The concatenative method relies on having a representative set of recordings that contain all the speech sounds of a language in a typical range of prosodic and phonetic contexts. The 'art' of the system lies in being able to construct an index of individual speech segments from which to select ideal tokens for concatenation. These tokens must be chosen so that they concatenate smoothly and at the same time carry the desired intonation and voice qualities. The acoustic samples are selected from the database for concatenation in novel sequences to make synthetic speech in the voice of the original speaker but with text, phrasing, and even language being freely manipulable.

CHATR was the first to use this method and for a time was considered the leading technique for creating natural-sounding synthetic speech. Licenses were bought by AT&T Bell Labs and NTT (the Japanese equivalent) among others, and subsequent versions of the system can be heard in many everyday applications providing voice-based services around the world.

This archive preserves the original CHATR voices and the samples that made the system famous at the time. Perhaps the main surprise is the small size of the archive given its coverage; the whole dataset being less than 500 MB in size.

The Japanese tour directory contains (for example):
11.html: CHATR Q & A
12.html: CHATR Testsuko's Opinion
13.html: CHATR Testsuko's Samples
31.html: Other Synthesis Systems : ATR NUU_TALK)
03.html: CHATR Examples
03_01.html: Examples (Multilingual Synthesis)
03_01_01.html: Examples (English)
03_01_02.html: Examples (Japanese)
03_01_03.html: Examples (Korean)
03_02.html: Examples (How a Speech Wave is made)
03_02_02.html: Examples (How a Speech ... Area Map Demo)
03_03.html: Examples (Selecting The Best Units)
03_04.html: Examples (Copy Synthesis)
03_05.html: Examples (Focus Shift)
03_06.html: CHATR Laughter
korean.html: CHATR Examples (Child Synthesis)

The e_tour focusses more on illustrating the diversity of voices that can be obtained by the system. It starts with the voice of Alan Alda, the anchor of a Scientific American television programme featuring the use of concatenative synthesis (in conjunction with image processing performed elsewhere) to produce a "Virtual Alan". His voice was made from a sample of 20-minutes that had been recorded by his studio as a challenge to the system. It also shows

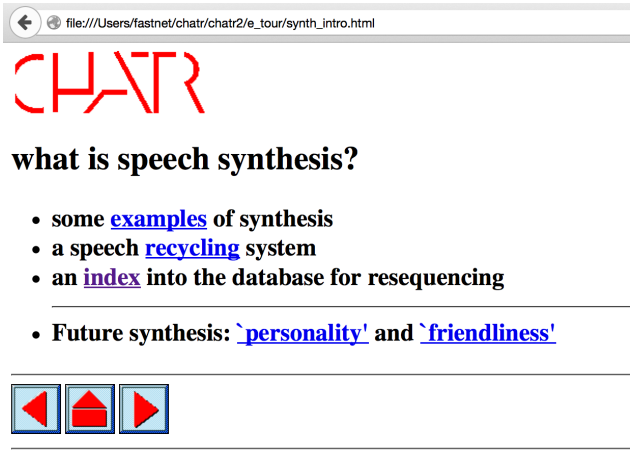

Figure 2: A sample page showing navigation links from the original web-site. The layout might be recognised as typical of PowerPoint or OHP slide presentations of the time.

how the voice of Kuroyanagi Tetsuko (a famous Japanese television personality) can be used to speak in both English and Korean as well as her native Japanese. Her voice was 'purchased' in the form of a cassette book, having two sides of slightly less than 30-minutes each. The recordings were digitised, indexed, and a database of her famous voice thereby made available for synthesis (with permission).

It also shows how children's voices can make very high-quality synthesis (Natsuki-chan was six years old, and her brother Yuto-kun then only four). The children first talk like themselves, but saying novel content, then they speak in (Japanese-accented) Korean, describing technical details of the synthesis system, and then produce some remarkably adult-sounding speech samples of a complexity that would not likely come from the mouths of such young voices.

The *e_tour* subdirectory contains:
ack.html: CHATR People
alda.html: CHATR's Alan ALDA
att_pr.html: CHATR Press Release
cd_synth.html: CHATR (random-access synthesis)
copy_synth.html: CHATR Examples (Copy Synthesis)
english.html: CHATR voices (English)
fkt_korean.html: TV Program on NHK
fkt_ml.html: Multi-lingual Tetsuko
fkt_samples.html: CHATR Testsuko's Samples
kaNsoo.html: CHATR Testsuko's Opinion
focus.html: CHATR Examples (Focus Shift)
fyo_korean.html: Chatr Ohta's Korean
german.html: CHATR's German synthesis)
girl.html: CHATR's newest girlfriend
japanese.html: CHATR Examples (Japanese)
kids_korean.html: CHATR Examples (Child Synthesis)
kids_ml.html: CHATR Examples (Child Synthesis)
korean.html: CHATR Examples (Korean)
larger_db.html: CHATR larger databases wanted
mks_ml.html: multi-lingual chatr (mks)
nyt.html: CHATR database example
qa.html: CHATR Q & A
sig_proc.html: (signal processing not yet!)
sig_proc0.html: (signal processing yes or no?)

4. Multilingual

Being a product of the Interpreting Telephony Research Labs (ATR), the multilingualism of the system was considered to be of great importance. Target foreign languages at the time included German, Korean, Chinese, and English, among others. The Japanese and English voices were usually recorded in a sound-treated studio using prompt sentences in each language, but the German database was taken (with permission) from the Kiel Corpus of Spontaneous Speech and synthesised samples include the voice of Professor Klaus Kohler, for example. The Chinese tonal variations were covered by use of PinYin transcriptions which include the tone as part of the phonetic label for each syllable. The Chinese samples were judged as good quality by natives at the time, but when asked which dialect they best represented, no clear answer could be found.

5. Behind the Scenes

An additional resource in this archive, though not yet open to public view is the code base.
All source code used to generate the samples is being preserved in both Windows (XP) and UNIX versions. Mac was in those days a specialist machine for artists and musicians. The user-interfface needs tcl/tk (version 3.0) to be installed but a command-line interface for converting text to speech in the various voices is sufficient if the standard GUI is no longer compatible with newer operating systems.

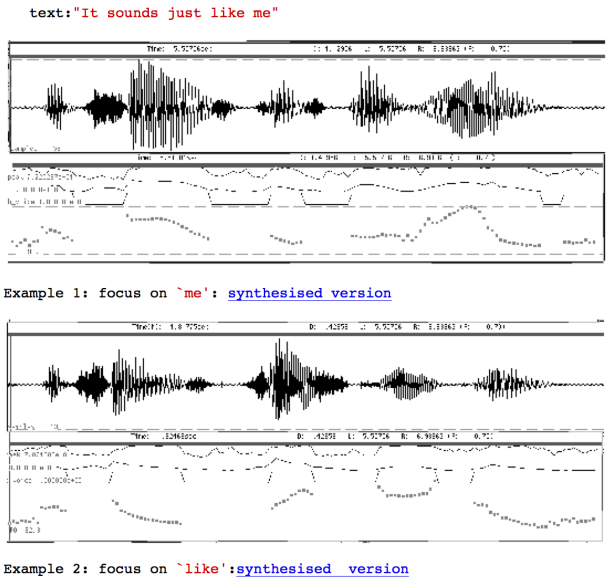

Figure 3: Part of a page created on request for the New York Times (e_tour/nyt.html), showing how the same text ("it sounds just like me") can be synthesised by this concatenative method without signal processing but with two different meanings: one stressing similarity "just like", and one stressing the target person "me". This page also contains screenshots of the individual waveform segments that were used in making up each sound and gives details of the types of context from which they were extracted.

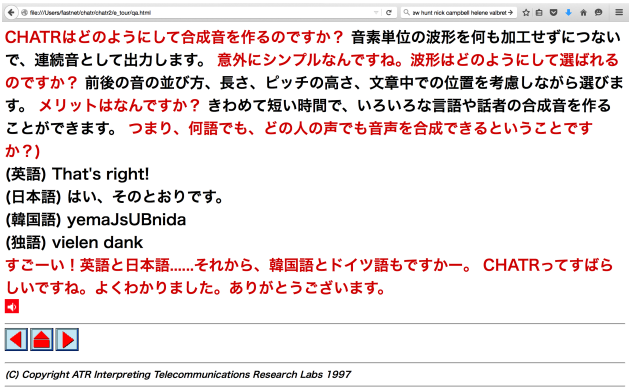

Figure 4: Perhaps the most often-visited page, Q&A using the voice of Kuroyanagi Tetsuko (a famous television personality) asking how the system works and being given explanatory responses in four languages. The text in red is the answer to each of her questions spoken in our favourite female voice (a Japanese/American bilingual)

The original tcl/tk installers are of course preserved as part of the archive and can be made available with the original source code to interested researchers wishing to work with the system either to produce new utterances in the original voices or to improve the original code for use with newly-created voice databases. For those with an interest in the code, the following operations are maintained:

General
GetChatrBuildInfo Return build information
SetDBPath Set the directory where speech databases can be found
Prosody
SetDurationStretch [mult] Set duration multiplier to mult
SetPauseDurations [b0 b1 b2 b3 b4 b5] Set pause durations for each break index
ProsodySynth [romaji] Generate a vector of targets from Japanese romaji input
ProsodySynthEnglish [text] Generate a vector of targets from normalized English text
SetDictionaryPath [path] Set the path where (English) dictionaries can be found
ProsodyGetTargs Return a vector of targets, corresponding to the last call to ProsodySynth or ProsodySynthEnglish
ProsodyPhraseTree Return a text representation of the phrase tree corresponding to the last call to ProsodySynth or ProsodySynthEnglish
ProsodySpeaker [speaker] Load speaker-specific parameters
SetHeadTailPause [head tail] Add pauses of given length to the head and tail of the target vector
UDB
UDBSetParams [params] Set global search parameters
UDBSearch [targvec Search the unit index for units corresponding to the target vector
UDBLoadSpeaker [speaker] Load a unit database index
UDBGetParams Return a list of global search parameters with their current values
UDBGetUnits Return a vector of unit labels corresponding

to the last call to UDBSearch
UDBGetUnitSegs Return a vector of unit labels intermixed with their corresponding targets
UDBGetSelectionInfo Return a table of information comparing targets with selected units
Concat
ConcatPlay [unitvec] Concatenate the given units and play through system audio
ConcatPlayStereo [unitvec level balance] Concatenate the given units and play through system audio
ConcatSaveWave [unitvec outfile] Concatenate the given vector of units and save the result in RIFF format
ConcatSaveUlaw [unitvec outfile] Concatenate the given vector of units and save the result in ulaw format
ConcatPlayWave [speaker waveid] Extract a wave from its wavlib and play it through system audio
ConcatPlayUlaw [unitvec] Concatenate the given units and play the result through system audio
ConcatTimeLeft Return the estimated amount of time remaining to complete the currently playing waveform
ConcatSetLevel [level Sets output volume level

6. Waveform Databases

The following original voice databases are being preserved, but are not currently open for public inspection or use. They can be made available to interested researchers on request. Name (f=female, m=male + two-letter initials) and Size (in megabytes): f2b 101M, fac 150M, fan503 89M, fhs 179M, fmm 102M, fmp 96M, ftk 96M, ftn 69M, fyo 105M, gsw 17M, mht 89M, mjb 476M, and anlp 119M (an exception to the naming rule).
In light of current speech corpus sizes, and especially considering the prevailing multimedia corpora of several terabytes, it is of interest to note the small size of what were then considered very large voice databases, on the order of an hour of recordings each.

7. Contributing Personnel

CHATR is the result of contributions by many members of Dept 2 of the ATR Interpreting Telecommunications Research Laboratories, and was developed under the supervision of the author who is especially grateful to Alan Black and Patrick Davin for much of the system integration and core programming. The name CHATR was registered as a trademark on May 13th 1997, and the process is covered by a patent (Campbell & Hunt) registered in 1996 as "Speech Indexing for Re-Sequencing Synthesis".

- Contributing Researchers: Yoshinori Sagisaka, Norio Higuchi, Nick Campbell, Alan Black, Naoto Iwahashi, Nobuyoshi Kaiki, Helene Valbret, Paul Taylor, Paul Bagshaw, Andrew Hunt, Toshio Hirai, Ken Fujisawa, Wen Ding Mingyue Xiezang, Ashimura Kazuyuki, and many others too numerous to list in full here

- Guest researchers: Mary Beckman, Arman Maghbouleh, Colin Wightman, Hyan-hi Lee, Gregor Moehler, J.J.Venditti, Christophe d'Allessandro

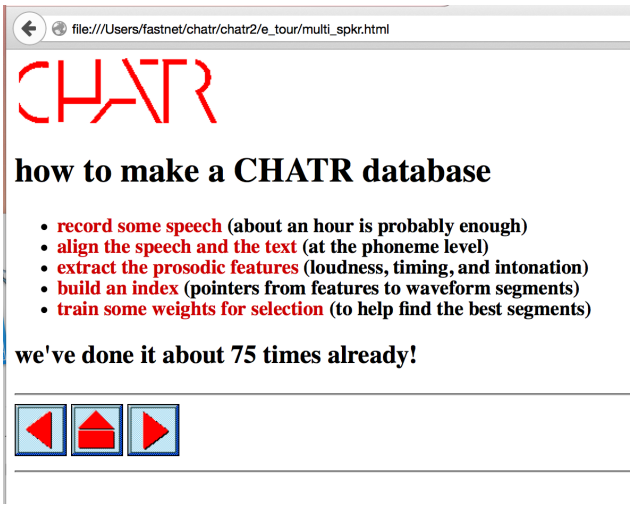

Figure 5: A simple (and very early) explanation of the concatenative synthesis process — from the archive

- Contract Programmers: Toshinori Satoh, Masahiro Nishimura, Yoshinori Mitijiri, Ken Shimomura, Shoushichirou Yonehara, and especially Patrick Davin

- Database Labellers: Yoko Ohta, Zai Hee Son, Kyoko Shimoda, Rachael Serrell, Sin-hwa Kang, Antonia Trueman, Chiyo Sakai

- Student Interns: George Hwang, Christian Lelong, Jean-Christophe l'Heriteau, Tony Hebert, Sebastian Voyard, Caren Brinkmann, Krisitna Striegnitz

- Technical Writers: Alan Black, Martyn Weeks, Hisako Satoh, Patrick Davin

- Project Leader: Nick Campbell

Acknowledgements

This archiving work is supported by the Science Foundation Ireland (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin. The tcd-fastnet.com site is supported under SFI grants 09/IN.1/I2631 and 07/SK/I1218.

References

[1] Campbell, N. "CHATR: A High-Definition Speech Re-sequencing System", in Proc ASA/ASJ Joint Meeting, Hawaii, 1996.
[2] Campbell, N., "Prosody and the selection of units for concatenation synthesis", pp 61-64 in Proc ESCA/IEEE 2nd w/s on Speech Synthesis, Mohonk, N.Y. 1994.