TTS IN EFL CALL – SOME PEDAGOGICAL CONSIDERATIONS

by **Włodzimierz Sobkowiak**

School of English, Adam Mickiewicz University

Poznan, Poland

sobkow@amu.edu.pl

Abstract

Rule-based Text-to-Speech synthesis (TTS) is discussed from the point of view of English as a Foreign Language (EFL) Computer-Assisted Language Learning (CALL). The perspective is pedagogical rather than technological. Some didactically salient characteristics of TTS are considered, such as (a) its feasibility as a pronunciation model, (b) control afforded over accentual and phonostylistic variation of speech, (c) the prospects of multimodal synthesis ('talking heads'). Some internet website addresses featuring TTS information, products and demos are provided.

1. Introduction

CALL is different things to different people. To some it is the use of text editors in the process of writing a homework assignment. To others it is surfing the internet for language tasks and exercises. To others yet it is when they stick that EFL CD in the drive and start on lesson number five. To learners it brings the welcome air of novelty to break the boredom of the language classroom. To teachers it means that extra effort and stress of preparing and running the technology-assisted lesson, with Murphy spitefully poised to spoil the carefully constructed scenario. To administrators CALL is that expensive foible of teacher X, forever in need of new equipment, software and servicing. To developers it is yet another opportunity to build in one of the newest hard- or software gimmicks to get the edge over competition. To parents it is sometimes the only motivation to finally give in and buy that idolatrous icon of modernity – the computer. To enthusiasts CALL is the promised land of language teaching and learning, with learners acquiring language knowledge and skills effortlessly from the machine, and the teacher walking benignly among workstations and laptops, offering help and guidance when needed. To technophobes CALL is a monster which can expose their ignorance and inadequacy, which threatens their (human/ist) ego, and which will ultimately destroy language learning as we know it today, transforming language teachers into beggars, and learners into cyborgs.

In the middle of all this commotion and controversy, CALL has now secured itself a safe position in all areas and on all levels of education, and it is certainly one of the most vibrant themes of didactic reflection and research at the beginning of the third millennium.

Some of its thrust comes from the tempestuous development of computer technology; some from the concurrent, but seemingly independent, changes in FLT methodology over the last decade or so – changes away from the pure communicative paradigm, in the direction of more form-oriented teaching and learning (e.g. Doughty & Williams 1998, or Ellis 2001). It is linguistic form, after all, that computers can (so far) manipulate much better than meaning or pragmatics.

In what follows I will consider the actual and potential impact of one of the cutting-edge computer technologies on one area of EFL which has traditionally been the most form-oriented of all – pronunciation. In particular, I will look at Text-to-Speech (TTS) synthesis as part of Human Language Technologies (HLT) or Natural Language Processing (NLP). The discussion will, however, be rather low-tech, with no reference to Hidden Markov Models (HMMs), Fast Fourier Transforms (FFTs), Dynamic Time Warping and the like. For these, the reader is referred to some introductory sources, such as Cole et al. 1996, Hovy at al. 1999 or Granström, House & Karlsson 2002. Instead, the perspective will be that of a (Polish) phonetics teacher and materials developer.

2. TTS in EFL CALL

There are, fundamentally, two types of speech synthesis: by concatenation, whereby previously recorded human speech is segmented and recombined, and by rule, whereby no previous human recording is necessary in any form, but rather a complex set of rules derives phonemic representations from spelling, and then generates the segmental and suprasegmental acoustics. While the process is rather more involved than what the preceding sentence might suggest, the TTS synthesis by rule has now overcome most teething problems and reached human-like quality (cf. e.g. Dutoit 1997 and 1999). The high-end TTS engines are rather expensive, and research to improve especially the prosodic properties of synthesized speech is still under way, but the technology is now reaching the stage where it can be applied to CALL.

2.1. Synthesized speech as a FLT pronunciation model

TTS synthesis has been in existence for a few decades now, mostly in applications for use by professionals and visually impaired people. The general public has first come into contact with 'robotic' speech in a variety of telephone information systems, such as train or air timetables. The two quality criteria proposed for such systems have been intelligibility and naturalness (d'Alessandro & Liénard 1996), rather than phonetic correctness with reference to some FLT norm. Thanks to both technological developments (processor speed, cheap memory, better DA converters) and linguistic research, all three criteria have now been reached. Not only is (top-quality) synthesized speech intelligible and natural (click here for a few short demos: http://elvis.naturalvoices.com/demos/), but it can also actually function as a model of pronunciation. For example, *Filoglossia*, a CALL package with (modern) Greek as a foreign language, already employs TTS synthesis: http://www.ilsp.gr/filoglossia_plus_eng.html, and WordPilot (http://www.compulang.com), also has this feature.

This creates a completely new situation in pronunciation-oriented CALL. So far, most of the CALL CD-ROMs (and most of the internet bandwidth for on-line speech-enabled courses and applications) was taken by audio (and video) recordings of speech in a variety of formats, including the space-saving mp3. It was believed that only good quality pre-recorded native speech can be profitably used in an educational CALL setting. But if the trick can be done with a TTS algorithm of a few megabytes, the freed space (bandwidth) can be used for other software, including e.g. artificial intelligence routines, which are crucial in sustaining a continued dialogue between the machine and the human. And, naturally, speech synthesis is by far less expensive than recording a team of highly trained human speakers.

The space and bandwidth savings are rather trivial technicalities, of course, from the point of view of FLT. Even if some CALL objectives and functionalities had to be compromised in the past due to space restrictions, the advent of DVD and wideband internet is guaranteed to change this facet of CALL completely. This has already happened in the US, and will eventually come to Poland as well.

2.2. Accent control

But there are more pedagogically interesting ramifications of TTS. One of them is the degree of control which the developer has over synthesized speech. Practically all speech variables can be manipulated at will. Take accent, for example. EFL CALL packages and electronic dictionaries for learners have so far catered for at most a few selected accents of English, usually the British RP on the one hand, and the American GA on the other. Some CALL programs would also include texts spoken in other accents, e.g. Australian, because learners' acquaintance with a variety of 'Englishes' has been at a premium in the communicative approach to EFL teaching and learning. But that was the limit of what could be achieved with pre-recorded speech. The learner could not listen to the same text spoken in different varieties of English as this would require wholesale duplication of recordings, and this was (and is) not feasible for a number of reasons. A well-tuned TTS synthesis system will allow the learner to pick the accent at will because – to the extent that accentual variation is rule-governed – the developer will be able to program the salient variables, such as, e.g. vowel quality, rhotacism, flapping, and the like for American English. The simple TTS synthesis plug-ins available for free off the web include the British and American accents as a matter of course (e.g. ReadPlease: www.readplease.com).

2.3. Phonostylistics

Nor is this the end of programmable phonetic variation, of course. Consider phonostylistics, i.e. variation related to the tempo and style (roughly casualness) of speaking. While accents are usually rather sharply categorized into their respective pigeon-holes, at least for pedagogical purposes in the EFL setting, the phonostylistic variation is evidently a continuum with no clear-cut boundaries, from overcorrect citation-form enunciation at one end, through reading and scripted speaking, to highly reduced allegro speech at the other end. In real life all this variation is heavily context-dependent, of course, with subject matter, situation, speakers/listeners, and a host of other factors all playing their roles. All this colourful phonostylistic kaleidoscope was of necessity mapped onto a grayscale of rather formal speech, with each text frozen forever in one stylistic rendition given to it by the recorded speaker. And yet, as is well known, understanding casual fast speech in naturalistic native settings is among the hardest tasks which the learner of EFL pronunciation must face.

Style-aware TTS synthesis could go a long way to help learners in their endeavours. The phonetic exponents of phonic styles in English are (at least partly) rule-governed and reasonably well understood: vowel reduction to schwa, schwa deletion, sonorant syllabicity, palatal coalescence, alveolar assimilations and elisions, for example, are all phonostylistically sensitive, and have gathered a sizeable bibliography. It is, I believe, mostly a matter of time for the TTS engines to be equipped with appropriate phonostylistic routines and algorithms. These will not only contribute to the overall impression of naturalness of synthesized speech, but will actually support a variety of phonostylistics-oriented tasks and exercises in CALL packages. Even (learners') electronic dictionaries would benefit, for – although today they only speak headwords rather than definitions or example sentences – even single lexical items spoken in isolation can be phonostylistically more or less appropriate: those marked as informal, slang or taboo are sometimes paradoxically pronounced in an incompatibly high style by the bored list-reading speaker.

To the extent that phonetic variation can be coded orthographically, tweaking orthography can crudely simulate a number of phonemic and allophonic processes, such as elisions or assimilations. For example, entering 'tem players' will TTS-synthesize a bilabially assimilated nasal alveolar, while 'coat' will obviously generate a Polglish-devoiced pronunciation of 'code'. There is ample space for experimentation here, but care should be taken as various TTS engines will react differently to nonce strings (mostly depending on the availability and structure of the built-in lexicon). For the sake of my T6 conference (http://www.ictconference.gliwice.pl/) presentation in Gliwice I attempted a rather crude approximation to the in-spe subtle control over phonostylistic aspects of English speech. I uploaded a short text for synthesis with the Festival TTS engine on the Bell Labs web page (http://www.bell-labs.com/projects/tts/; this service is now discontinued, but see ATT webpage at the URL address specified below). One version of the text had orthographic alterations meant to simulate some phonostylistic phenomena: final alveolar stop deletion, affricate weakening, nasal 'bleaching'. The text which I uploaded was (without brackets): "My name is Radek. I welcome all presen[] in hall C-1 at professor Sobkoviak's lec[s]ure. My task is convincin[] you abo[w] the high level of Bell speech synthesis". The effect (http://elex.amu.edu.pl/~sobkow/tts/RadekFastFestival.wav) is best audially compared to the unaltered version (http://elex.amu.edu.pl/~sobkow/tts/RadekFestival.wav): "My name is Radek. I welcome all present in hall C-1 at professor Sobkoviak's lecture. My task is convincing you about the high level of Bell speech synthesis". Some 'casual speech' effects are of course more convincing than others, but notice that all were achieved by a rather primitive method of orthographic manipulation. With full control at the level of the TTS engine, real phonostylistics can easily be attempted.

2.4. Foreign accent simulation

The accentual, dialectal and phonostylistic variation within the native norm does not exhaust the pedagogically useful control possibilities in TTS synthesis. It would be technically rather easy to simulate a foreign accent, for example to better demonstrate to the learner the areas which need improvement (e.g. final devoicing in Polish English). This is something which few native speakers would be capable of, and while it is not unimaginable to employ non-natives for the job, it would clearly verge on impossible to be able to control just the wanted phonetic variables to the exclusion of others. Expert non-native phoneticians could be resorted to for a better command of the phonetic interlinguistic intricacies, but this would again be troublesome for other reasons. The precise phonetic control afforded by a TTS system can hardly be improved.

Simulating a foreign accent of English by computer for didactic purposes is not a new idea. In 1997 Hyouk-Keun Kim created his *Korean Accented English Pronunciation Simulator* (http://odin.prohosting.com/hkkim/cgi-bin/kaeps/kaeps_home.htm), rightly noticing that "Most adult ESL/EFL learners [...] do not recognize the problems of their English pronunciation", and that it might be a good idea to demonstrate these under computer control. Eventually a rule-based KAEPS system was set up, simulating "three types of English pronunciations in the IPA symbols: 1) a phoneme-based English pronunciation, 2) a desirable allophone-based American English pronunciation, and 3) a possible Korean accented English pronunciation". While Kim's system has never advanced beyond accented graphemic (i.e. IPA) representation, it would be easy enough to attach the IPA-to-speech engine to it. After all, most TTS systems use phonetic transcription at some stage of the synthesis process.

Notice that, like in the case of phonostylistic variation, L1-accented English speech is really a continuum of interlanguage, roughly correlated with the traditional dimension of proficiency. To take a Polglish example: initially Polish learners of English will tend to force the English vowel system into the Procrustean bed of their native 6-point system. At this stage, all of English /ɑː/, /æ/, /ʌ/ tend to fuse into Polish /a/. Later /æ/ tends to emancipate itself and the back and central vowels take their respective positions, with /ʌ/ the last to be properly interpolated between Polish /a/ and /e/. An L1-sensitive TTS system would be able to dynamically adjust its parameters to realistically simulate spoken Polglish at these various stages of proficiency. Needless to say, the relevant phonetic bibliography on the dynamics of English-targeting interlanguage speech development is by far more modest than that devoted to intra-English variation. This is particularly true of such minority L1's as Polish, unfortunately. The bottom line is, then, that much more fundamental research would be needed to feed into the creation of a Polglish speech simulator.

2.5. Non-human speech

If the argument is accepted that listening to variably L1-accented speech, produced and manipulated under computer control, might help learners first notice and then get rid of their accent, the logical extension is to take it one step further, like Keller & Zellner-Keller (2000a) did, noting that "speech synthesis allows [...] the creation of sound examples that could not be produced by a human being (e.g., speech with intonation, but no rhythm)". Demonstrations like this are located somewhere between teaching foreign pronunciation as one practical language skill (on a par with reading and writing, say) and teaching foreign phonetics as declarative knowledge. The latter is done as part of the so-called 'descriptive grammar' in the Polish academic philological setting. The recent rapprochement of these two kinds of courses, partly stimulated by the 'focus-on-form' movement in EFL mentioned earlier, makes the Kellers' idea quite attractive. Notice, incidentally, the peculiar paradox: TTS synthesis was initially developed (and is still perfected) to make artificial speech sound as human-like as possible, but for some didactic applications it is actually beneficial to create more 'robotic' speech.

2.6. Visual synthesis

The most advanced TTS systems now available go beyond the simple (?) unimodal audio synthesis, into the exciting world of face animation, featuring the so-called 'talking heads' or animated agents. One of the most successful applications of this cutting-edge computer technology to CALL has been the University of Colorado Center for Spoken Language Understanding (CSLU) "Baldi" project (http://cslr.colorado.edu/toolkit/main.html). In brief, it is an NLP environment focused on the use of TTS synthesis and Automatic Speech Recognition (ASR), enhanced with the animated face ("Baldi", Figure 1; Baldi has recently been superseded by other faces, but the essentials of the system remain) simulating phonetically realistic articulatory movements in real time.

Figure 1. "Baldi". http://elex.amu.edu.pl/~sobkow/tts/Baldi1.jpg

Visual object programming, speech spectrography and many other components are integrated in the Rapid Application Developer which makes it possible to create a simple dialogue schema in minutes, which can then be built into another application, such as CALL for example (see http://www.haskins.yale.edu/haskins/heads.html for a comprehensive interactive overview of many other 'talking head' projects).

What is most exciting in the package (which is free for educational purposes) is the novelty of using the animated face to enhance speech synthesis and make the spoken exchange more realistic. Baldi not only moves his lips and eyes to provide the much needed – especially in the context of learning a foreign language – visual information to aid intelligibility. It can also 'go transparent', exposing the realistically rendered inner articulators in full motion, down to the root of the tongue (see Figure 2).

Figure 2. Baldi gone transparent. http://elex.amu.edu.pl/~sobkow/tts/Baldi2.jpg

This is an incredible resource for pronunciation learners, of course: they can listen to natural (if synthesized) speech and see how it originates in the mouth. The head is quasi-3D; it can be rotated in all three dimensions with the mouse, and the amount of transparency can also be adjusted at will, the extreme leaving just the articulators on screen.

The CSLU toolkit, where Baldi lives, has so far been used mostly to assist speech and language therapy of native American children, but its application to EFL CALL (and other L1's – Baldi can be programmed for any language whatsoever) is just a matter of time. Also, it is enough to go to the movies nowadays to see the level of realism which animation of human-like synthetic actors has achieved (e.g. "Shreck" or "Lord of the Rings"; see also Thalmann & Thalmann 1990). In a few short years animated anthropomorphic agents will be used in CALL, which will be hard to tell apart from video-recorded real human speakers. One technical consequence of this will be – like with the TTS synthesis – that more CD space will be freed from the enormously memory-hungry current video files. It is much harder to predict learner reactions to (semi-intelligent) speaking and animated human-like agents acting as conversation partners in settings which are now only available in video conferencing. Learners may relate to these artificial personas to the extent which may be pedagogically relevant, with both its pros and cons.

3. TTS on-line demos

There are many players in the field of TTS synthesis, both academic and commercial. Most would offer some information about their research and products on the Internet, including passive and interactive demos. The latter are of course by far more exciting from the point of view accepted here, i.e. that of a EFL pronunciation teacher. While most demos would only allow to enter a rather short piece of orthographic text for synthesis, some measure of deliberate spelling manipulation is afforded (like I did above with Radek).

Without attempting a comparative description and analysis of the many TTS systems available on the Web, I will close this section with a few best-known URL addresses, and let the interested reader experiment on his/her own.

Some of the best-known TTS systems and their manufacturers:

- Centre for Speech Technology Research: http://www.cstr.ed.ac.uk/projects/festival/
- ScanSoft's RealSpeak: http://www.scansoft.com/realspeak/demo/
- Prosynth Project: http://www-users.york.ac.uk/~lang19/
- ATT Labs: http://www.research.att.com/projects/tts/
- Elan Speech: http://www.elantts.com/

4. Conclusions

Text-to-Speech Synthesis is of course only one branch of the NLP tree. In this text no attention was paid to Automatic Speech Recognition (ASR), Machine Translation (MT), multimodal man-machine communication, automatic information extraction (data mining) and summarization, language generation, multilingual resources, speaker/language identification, speech evaluation techniques, Artificial Intelligence (AI), and a number of other areas in one way or another concerned with natural language processing. All of them are potentially of interest to foreign language educators. Some have been recognized as such, and work is going on exploiting their potential, e.g. ASR. Some are still *in statu nascendi*, struggling with technological problems and lack or inadequacy of relevant linguistic research, e.g. AI. But it is a safe guess that sooner or later (most probably – sooner) all of these human language technologies (HLTs) will arrive at the door of Foreign Language Teaching, EFL in particular (and soonest), demanding to be accommodated. FL teachers should better be prepared, lest pedagogy be compromised for sheer technology.

References

d'Alessandro, C., Liénard, J.-S. (1996) "Synthetic speech generation". In R.Cole et al. (eds). Chapter 5.2.

Cole, R. et al. (eds) (1996) "Survey of the state of the art in human language technology". http://cslu.cse.ogi.edu/HLTsurvey; accessed 13.4.2003

Delcloque, P. (ed.) (2000) *Proceedings of InSTIL: Integrating Speech Technology in Learning*. University of Abertay Dundee, Scotland.

Doughty, C., Williams, J. (1998) *Focus on form in classroom second language acquisition*. Cambridge: Cambridge University Press.

Dutoit, T. (1997) *An introduction to text-to-speech synthesis*. Dordrecht: Kluwer Academic.

Dutoit, T. (1999) "A Short Introduction to Text-to-Speech Synthesis". http://tcts.fpms.ac.be/synthesis/introtts.html

Ellis, R. (ed.) (2001) *Form-focused instruction and second language learning*. Oxford: Blackwell.

Granström, B., House, D., Karlsson, I. (eds) (2002) *Multimodality in language and speech systems*. Dordrecht: Kluwer Academic.

Hovy, E. et al. (eds) (1999) *Multilingual information management: current levels and future abilities*. http://www.cs.cmu.edu/~ref/mlim/index.html

Keller, E., Zellner-Keller, B. (2000a) "Speech synthesis in language learning: challenges and opportunities". In P. Delcloque (ed.)

Keller, E., Zellner-Keller, B. (2000b) "New uses for speech synthesis". *The Phonetician*, 81, 35-40.

Thalmann, N., Thalmann, D. (1990) *Synthetic actors in computer-generated 3D films*. Tokyo: Springer Verlag.