PRONUNCIATION IN EFL CALL [1]

by Wlodzimierz Sobkowiak

Adam Mickiewicz University,

Poznan, Poland

sobkow@amu.edu.pl

Abstract

State of the art in pronunciation-oriented EFL CALL is reviewed from the pedagogical perspective. Discussion touches upon CALL flexibility, coverage, declarative vs. procedural knowledge, L1-sensitivity, multimedia employment and automatic speech recognition (ASR). Six different CALL programs are briefly evaluated from these points of view: *Fluency, Pronunciation Power, Connected Speech, Better Accent, ISLE* and *Tell Me More*. Future promises and challenges in speech-enabled EFL CALL are outlined, such as speech synthesis, multimodality in man-machine communication and (speech-to-speech) machine translation.

1. Introduction

All three terms appearing in the title of this text beg for definitions, explanations and discussion. They are in the mainstream of current foreign language pedagogy, on the one hand, and they are all multiply ambiguous, on the other. To illustrate the former claim, one would only need to mention the heated dispute on the standards of English-as-a-Foreign-Language (EFL) pronunciation teaching, which has been raging over Europe for some time now, and has found some reflexes in Poland as well [2]. Or the well-known fact that English is the most widely taught foreign language in the world, with views to become the one and only true *lingua franca* of the globe. Or the tempestuous development of computer-assisted (English) language learning (CALL) over the last decade or so. To illustrate the latter claim, it would be enough to point at the notorious fuzziness of 'foreign' in EFL (as opposed to 'second', for example), or that of 'learning' (in CALL). Even 'pronunciation' turns out to be definitionally problematic on certain levels of phonetic reflection, if only because it is not a simple synonym of 'phonetics'.

To start this discussion at this point, however, would be to jeopardize the main aim of the paper by getting swamped in layers upon layers of metalinguistic and methodological details and distinctions. And the main aim of the paper is, after all, not to conduct methodological discourse, but rather to present a critical snapshot of the current state of art at the interface of the three areas

listed in the title. This interface area itself is enormous and it tends to exhibit a breathtaking pace of innovation, mostly due to the rapid development of computer technology (some authors prefer ICT – Information and Communication Technology – but I will stick with the simpler term). In this situation, I can entertain no hope of ever managing to provide a comprehensive and completely up-to-date account of the whole area. The sampling I made is of necessity subjective and fragmentary. For example, mainly for reasons of space, I am not going to venture into the exciting world of on-line Internet pronunciation teaching and learning, even though distance education is among the most fashionable themes in current foreign language pedagogy. The discussion will thus be restricted to 'localized' CALL, which could also be called 'off-line' or CD-ROM- (or DVD-) based. Even so dramatically circumscribed, the area is still too large to treat representatively. Out of many potentially interesting issues I will select only a few. Out of hundreds of available CALL programs, I will present but a handful. Out of their many components and functionalities, I will concentrate on those which I regard as central to my theme.

The organization of the remainder of this text is as follows: first I will discuss some key issues in EFL CALL pronunciation, then some relevant software will be presented and briefly analyzed from the point of view of the preceding discussion, finally a rather informal glimpse of the future will close the paper.

2. Computer-assisted pronunciation teaching and learning

2.1. Flexibility

CALL researchers have successfully argued (e.g. Kaliski 1992, Warschauer 1996, Warschauer & Healey 1998, Kern & Warschauer 2000) that one of the main strengths of CALL is its didactic flexibility. Unlike some other educational technologies which have been implemented in schools over the ages (from blackboards to video, say), computers will fit any didactic approach, method or technique, if used skillfully. Grammar-translation supporters may use them to expedite translation practice from L2 to L1, for example, with machine-translation software. Audio-lingualists will be able to control their students' structure drilling with the computer in much more sophisticated ways than they could in the traditional language laboratory. Cognitivists will sit their learners in front of adventure games, where they will have to navigate an unknown territory using their linguistic competence. Communicatively-minded teachers may pit learners against each other in a simulation game to make them negotiate meanings to reach agreed-upon goals.

This flexibility of CALL is true also on phonetic ground. Practically all multimedia programs presented below, though to varying degrees, can be accommodated into different pronunciation-teaching philosophies. Specifically, both the 'know-that' declarative knowledge component, and the 'know-how' procedural skill component of phonetic competence can be appealed to in various ways, for example through multiple-choice exercises and listen-repeat-compare tasks, respectively. Despite common belief, pronunciation-oriented CALL has not given up on the old techniques in favour of the razzle-dazzle display of vibrant hypermedia. Rather – quite wisely in

my opinion – it has accommodated the new technological achievements such as speech recognition, for example, into a versatile framework of structures and functionalities where each learner and teacher can find something to fit his needs and preferences.

2.2. Coverage

An issue related to CALL's flexibility is its coverage. The classic core of pronunciation training in the traditional syllabus includes segments, suprasegmentals, fast speech phenomena, grapho-phonemics, accentual variation. On the level of particular textbooks, courses and materials there is enormous variation, of course, both in choice and priority of these elements. Communicative language learning, for example, brought with it the preference for prosody in pronunciation teaching, with some courses actually beginning from rhythm, stress, juncture and intonation. This is now changing, with the advent of the post-communicative era in foreign language teaching. Regardless of fashions and vacillations, however, the canon is reasonably well defined.

Contemporary pronunciation-oriented CALL is able to deliver instruction in all those canonical areas. There are programs concentrating mostly on individual sounds of English, as well as those which cater predominantly for suprasegmentals. Some would specifically target natural (fast) speech, while others would proudly (and politically-correctly) offer different accents from speakers of different skin colour. This is not to say that a particular piece of software will necessarily include the full phonetics syllabus. Unlike with the methodological flexibility of section 2.1., coverage of phonetic substance is of necessity much more 'hard-wired' in the structure of the package. It will be obvious from the short software presentations below that programs tend to specialize in certain sub-areas of the pronunciation syllabus. What is crucial, however, is that there is now no technological obstacle to using CALL in any of the canonical components.

2.3. Declarative versus procedural knowledge

This dichotomy was mentioned above in 2.1. It appears to be among the most fundamental distinctions in all foreign language teaching, including teaching pronunciation. It captures the intuitively rather obvious truth that in order to do anything one must have – in varying proportions depending on the actual activity – both the 'theoretical' and 'practical' competence. Unlike syntax and vocabulary, pronunciation in a foreign language has traditionally been regarded as the exclusive province of the latter; hence murderous drilling in the audio-lingual method and little explicit teaching in the cognitive-communicative era. The pendulum now appears to be swinging in the other direction, so that the declarative, explicit, 'know-that' meta-competence is back in the picture, with researchers trying to feed it into the process of phonetic proficiency building (see, e.g. Dziubalska-Kołaczyk 2002). In the academic context of EFL, for example, this means that the so-called 'descriptive grammar' of English should be well integrated with the practical phonetics syllabus, so that students practicing, say, the intricacies of English

obstruent voicing could fall back on their knowledge concerning laryngeal excitation source-filter models, as well as 'external' evidence from their L1 interference, speech errors, speech play, and the like.

CALL supports both types of knowledge. Most multimedia pronunciation programs are not content to provide the learner with ample opportunities to use his articulators, whether for simply recording utterances or for actual simulated dialogues. There is usually also explicit instruction concerning such matters as correct articulation and voicing, keeping the right rhythm, varying the pitch for intonation, using appropriate lexical and sentential stress patterns, and the like. This instruction may appear in many different forms in the program: as mini-lectures, glossaries, multimedia presentations, help files, task prompts, error messages, and many others. Some packages offer manual-like functionalities which can be used more or less like traditional textbooks, complete with comprehension questions and suggestions on further reading.

2.4. L1 sensitivity

Where there is still a lot of room for improvement is how CALL relates to the native tongue of the user. The sad truth is that in very many cases it simply does not. There appear to be two main reasons for this state of affairs, one commercial, the other linguistic. The former has to do with cost-effectiveness mostly: to prepare a large CALL package with all the currently expected multimedia bells and whistles is an extremely expensive undertaking, much more so than, say, a traditional course with manuals, exercise books and audio tapes. The large investment will only pay off if the package can be used on a global scale with all thinkable L1 learners. Investing in a multi-CD EFL course for Poles only, for example, is hardly commercially viable. Exceptions to this rule are small programs made locally or localized versions of the leading packages made in the West. Even these, however, seldom go beyond simply translating the metalanguage and fitting local translations to the existing monolingual built-in dictionary.

This appears to be due to the other reason mentioned above. While we may know a lot about L1 transfer and interference on the theoretical level, there are huge lacunae of knowledge in the actual application of this information in the speech-enabled CALL setting. For example, the CALL craze of the last few years – speech recognition (see below) – has hardly reached a stage where it would have a viable model of the learner with a given L1, hence a particular interlanguage. The technique is hastily transferred from native speaker applications such as dictation or dialoguing expert systems into the world of EFL with little recognition of the need to make it sensitive to non-native speech. Thus, what one often observes is either disastrous recognition results with highly demotivating end-effects, or the anything-goes principle where any learner input is happily accepted. Both these extremes are thoroughly a-pedagogical, of course, as noticed by many researchers (e.g. Chen 2001). I can but agree with Ehsani and Knodt (1998:56) at this juncture that "one of the most needed resources for developing open response conversational CALL applications is large corpora of non-native transcribed speech data, of both read and conversational speech".

2.5. The growth of multimedia

'Multimedia' is one of the modern buzz-words, on a par with CD-ROM, SMS, hypertext, DVD, mp3, video-clip, and dozens of others. Few young people in the developed countries would be completely ignorant of it, and most would agree that the term has positive connotations with novelty, movement, sound, colour, fun, etc. It is these connotations that are exploited in the contemporary saturation of CALL with multimedia. CALL, after all, is supposed to be edutainment, it must motivate, it must attract. And what better attraction to a young mind can there be than a full-colour video with good quality sound? Most current CALL packages are built on this premise (one reason why they are expensive to make).

Multimedia has grown gradually. First, (still) pictures were added to sheer text, then sound of initially rather poor quality, then simple animations, finally video. Pronunciation-oriented CALL jumped on each band-wagon soon after they appeared. In the age of CGAs and Hercules graphics no phonetic transcription (or accented letters, for that matter) could be shown on screen, so simplified systems had to be used[3]. With the first graphics showing articulators in cross-section, vowel diagrams or lip shapes added to text CALL resembled good old pronunciation manuals. With the advent of animation and audio the true era of multimedia began and pronunciation-CALL finally got its added value. Sounds could now be illustrated in various media: in transcription, articulatory diagram or recording. Finally, with the improvements of sound cards, sound recording algorithms, processor speed and memory, sounds could also be visualized as waveforms or (for the more intrepid learners) as spectrograms.

It is this last idea that has become enormously popular among the designers of CALL software and among the less critical users. It is usually implemented in the context of the listen-record-compare task: the learner listens to the model (native-speaker) recording, records his own rendition of the text trying to mimic the original closely, finally compares the two recordings along both channels: aural and visual (example screen-dumps appear below, e.g. Figure 12). The former – aural comparison – is the traditional method used since the very inception of foreign language learning. The latter is new (motivating element!): the eye is supposed to guide the tongue, to put it metaphorically. While the technique obviously has the required commercial potential, there are serious pedagogical objections to its use (see e.g. Ehsani & Knodt 1998). To enumerate them briefly:

(a) no two recordings, even of the same person, are exactly acoustically alike,

(b) no instruction is normally provided on how to align the two waveforms for best result,

(c) speech tempo and loudness will interfere with the correct reading of the waveform information,

(d) considerable acoustic knowledge and skill are needed to be able to benefit from comparing waveforms,

(e) the technique is very sensitive to hardware quality.

This is hardly the end of multimedia development, of course. New ideas and technologies hit the news lines weekly. Some of them have serious ramifications for the future of pronunciation-aware CALL. These will be briefly presented at the end of this paper in the 'science fiction' section.

2.6. Automatic Speech Recognition (ASR)

It was at the beginning of the last decade of the 20th century that computer hardware reached a stage where it could support speech recognition in real time. This was of course an enormous breakthrough in computer-human communication, which had so far been multimedial in the output, but monomedial (keyboard) in the input (see Aist 1999 for an overview and literature). As mentioned above, ASR was immediately implemented in a number of consumer applications, starting with those where single-word input was sufficient, and gradually spreading to other functionalities, such as dictation, for example. To fully appreciate how complex computer speech recognition is one would have to have a large aside here explaining the intricacies of Hidden Markov Models (HMMs), Fast Fourier Transforms (FFTs), Dynamic Time Warping, and a bunch of other algorithms verging on higher math. The reader is referred to some introductory sources, such as Deroo, Bernstein & Franco 1996 or Cole & Zue 1997. Here, we will briefly concentrate on CALL applications of ASR.

Notice first that the term itself may be a misnomer if it is applied to such functionalities as the mentioned waveform display, for example, or the facility which tries to automatically assess the acoustic fit of the learner-recorded speech with the model (as some of the programs analyzed in section 4. do). No true recognition is required in either process, rather acoustic pattern matching, which is technologically much simpler. For fully-fledged ASR to take place, the software must meaningfully react to the content of the input speech. This meaningful reaction may simply be visualizing speech as computer-readable text on screen, or adjusting the flow of the simulated dialogue according to learner input, for example. This is the current stage which ASR has reached in (the more advanced) CALL, as illustrated below (see *Tell Me More* in 3.6.). Notice that the dialogue is simulated, i.e. the learner can control the flow of the conversation within limits imposed by the ASR algorithm, which will only allow the choice of a few communicatively relevant responses to the computer's recording (a low-perplexity closed-response task for the ASR engine). With no true AI (Artificial Intelligence) in view, the simulation can proceed no further: any attempt to enter into natural conversation with the machine quickly ends in linguistic-communicative disaster, as clearly illustrated by the many 'chatterbots', i.e. chatting robot-like agents on the web (see selection of links on www.chatterbot.tk, for example). ASU (Automatic Speech Understanding), unlike ASR, is still a long way off.

With all these caveats, ASR did give CALL an added measure of face validity. To be able to speak to the machine in the foreign language and have it react in meaningful ways is certainly exciting to most learners, especially to the new generation of children, who take the 'traditional' modes of keyboard communication for granted. Also the speech assessment routines now built into some ASR-equipped CALL can initially be quite thrilling. From them the learner will get

the extra metalinguistic feedback on his pronunciation, as if from the teacher. This may initially motivate the learner to actually try harder and pronounce the given sentence again, hoping to push the match indicator to an even higher level. This repetition is no bad thing, of course, as far as it goes. Learners quickly discover, however, that the mechanism can hardly offer robust and detailed evaluation and guidance, as a teacher would: all that can be expected is a global yes/no measure of phonetic achievement, with completely mysterious evaluation criteria and no explanation whatsoever[4]. With the user-customized acceptance threshold set to low, this functionality will accept virtually any spoken input and grade it as good; with high settings even native speakers will have problems getting satisfactory notes. Thus, somewhat analogically to the visual feedback routines discussed earlier, the technique is highly pedagogically questionable at this stage of ASR development (see Chen 2001 for similar conclusions).

3. Pronunciation-oriented CALL software: a sample

Just like the selection of issues in EFL CALL pronunciation in section 2 of this paper, so the choice of software for presentation and analysis is unavoidably fragmentary and subjective. With the line of CALL merchandise now running into hundreds, it could hardly be otherwise. Few comprehensive overviews of (EFL) CALL products exist, also because it is almost impossible to keep pace with the growth of the field: new packages hit the market every week. Finally, serious CALL journals, can only carry a few reviews in every issue, thus sampling a tiny proportion of the whole market. In this situation, the only feasible resource for CALL software information, analysis and advice is the Internet. And indeed, there is CALL info galore on the web, with the characteristically unavoidable disadvantage of uncertain reliability and expertise, volatility, advertising hype, selectiveness, repetition and dispersal. With the practical unavailability of other sources, however, the web remains the best place to go for CALL software research.

In what follows I will briefly present six pronunciation-oriented EFL CALL programs, relating the discussion to the issues discussed in section 2. The programs are listed in the table.

As can be seen from the table, despite the tiny size of the sample, the packages do represent a certain variety along the four dimensions of (a) visual and (b) corrective feedback, (c) transcription use, and (d) general didactic focus. Each one has some www presence in the form of their dedicated websites. They are enjoying a certain amount of critical reviewing attention, as well. The following is not to be construed as actual reviews of the programs, however, as explained earlier.

3.1. *Fluency*

According to the blurb on the Fluency's website, "*Fluency* is designed to let you speak, then give you feedback as to how you did – what to correct and how to correct it. Using state-of-the-art speech recognition technology, SPHINX from Carnegie Mellon University, this interactive software allows you to speak, to get corrections, to listen to yourself and a native speaker and try again, over and over, as many times as you want". The website is that of the Language Technologies Institute (http://www.lti.cs.cmu.edu/index.html), which does most of its research in machine translation and information retrieval, and *Fluency* appears to be an offshoot of that research. Going a little deeper into the LTI website, we will encounter a slightly more precise description of the *Fluency's* operation: "The system detects pronunciation errors, such as duration mistakes and incorrect phones, and offers visual and aural suggestions as to how to correct them".

As seen in the screenshot in Figure 1, the program's functionalities are rather modest, exactly like stated above. The sagittal cross-section of the vocal tract and the frontal lip view are now practically standard features of EFL (pronunciation-oriented) CALL software (sporting better graphics in most cases). The obscure non-IPA phonetic transcription system, on the other hand is definitely non-standard (even if used in practically all CMU Natural Language Processing (NLP) applications). The segment illustrated, i.e. AX is of course a schwa. Unlike most EFL CALL programs, *Fluency* does try to evaluate and correct the pronunciation of single segments, but the user would be hard-pressed to find out something substantial about the criteria the software is using in the process, so self-correction remains a hit-and-miss procedure.

There is a lot of the declarative-knowledge material in the program: the transcription, the diagrams, the corrective advice are all phrased in the notorious 'descriptive grammar' parlance. The speaking skill component is restricted to simply reading a reply in the simulated dialogue. The learner can also listen to the model utterance, of course, but this would hardly count as a communicative activity, say practicing listening comprehension, because there is effectively none. Neither is there any pretense of a communicative setting, of course, with the short dialogue highly stylized and unnatural.

3.2. *Pronunciation Power*

Pronunciation Power is marketed by English Computerized Learning Inc. (ECL), which, according to their website, "was founded in 1995 in Edmonton, Alberta, Canada [and] operates as a developer and distributor of professional high quality, interactive, multimedia ESL/EFL CALL materials". The package was developed, however, by Blackstone Multimedia Corporation, which is a privately held company also based in Edmonton. It is enough to have a look at their respective websites to appreciate that, unlike in the case of *Fluency*, the *Pronunciation Power 2* package is at the very core of the companies' business, reportedly "used and recommended by over 4000 universities, colleges, businesses and schools worldwide".

Pronunciation Power is designed for over a dozen L1s. It includes a variety of game-like activities and functionalities, such as (a) listen-record-compare, with single words, minimal pairs and full sentences, (b) listening discrimination, (c) vocal-tract cross-section and lip animations

(Figure 2), (d) waveform display and comparison (Figure 3), (e) STAIR exercises: stress, timing, articulation, intonation, rhythm, (f) a 40-page manual, (g) on-board illustrated audio dictionary of over 7,000 words with a variety of search keys. Thus, it is a large and fully professional package, entirely devoted to pronunciation training, enjoying a considerable commercial success.

With the multitude and variety of material and activities it offers, *Pronunciation Power* is suitable for pronunciation teaching and learning under different conditions and in different settings, thus eminently proving that flexibility is a great asset of CALL. Its coverage, in terms of the canonical pronunciation/phonetics syllabus is large, although fast-speech phenomena and accentual differences of English appear to be underrepresented. The declarative-procedural balance is much better than in *Fluency*, with a lot traditional textual exposition on the one hand (see Figure 4, for example, where heavily technical terminology is used to annotate the diagram), but also the extremely rich skill component containing varied activities for practicing both listening and speaking. In going through the demo of the package, which is freely dowloadable from its website, I did not notice any L1-sensitivity beyond the localization of the dictionary. This is, as discussed above, the sad norm in native-made EFL CALL. While the waveform matching technique is used to its limit (with the learner being able to drag-align his recording to better fit the model; see Figure 3), no attempt is made to automatically evaluate the learner's pronunciation. Considering the criticism I voiced earlier, this is probably to the good of the package. While the system of phonetic transcription used is thoroughly simplified compared to IPA, it is still generally IPA-ish, with ashes, schwas, engs, thetas, and the like[5]. ASR is used for most exercise types, so that there is relatively little keyboard input.

3.3. *Connected Speech*

Connected Speech was made by an Australian firm Protea Textware. Like *Pronunciation Power*, it is an integrated CALL package exhibiting professional design and programming, with noticeably less content, versatility and multimedia interaction, however. From the entry screen (see Figure 5), the learner can be taken to one of three proficiency levels (starting with lower intermediate) in one of the accentual versions, American, Australian or British English, placed on separate disks. There, he should first listen to a few minutes of video-recorded narrative monologue, whereupon he can enter a suite of tasks and exercises mostly focused on "the suprasegmental features of English, with mark up, recording, practice activities, tests and tutorials. It has speech recognition that gives specific feedback on the suprasegmental features of the learner's production" (from http://www.proteatextware.com.au/). As can be seen in Figure 6, separate components are dedicated to such phonetic areas as pause groups, stress, pitch change, linking, segmentals and syllables. More traditional exercises are also included, such as listening comprehension tests prompting the learner to fill in critical vocabulary items from the keyboard, or IPA training.

The program does appear to focus on connected speech, which makes it unique among those under scrutiny here, although the monologues are rather far from the native norm of "natural informal", even at the advanced level, the web blurb notwithstanding.

The pros of *Connected Speech* are: (a) the good balance between the declarative and procedural element, (b) the wide variety of voices, tempos and accents, (c) the skillful use of ASR on all phonetic levels, i.e. for segmentals, pitch, stress and duration, and (d) a well-designed, uncluttered and intuitive user interface. Among the cons one should certainly mention: (a) the rather disappointing use of video to record 'talking heads' only[6], (b) complete L1 insensitivity, and (c) the generally rather uninspired design of the tasks and exercises.

3.4. *Better Accent Tutor*

American-made *Better Accent Tutor* is "pronunciation training software based on instant audio-visual feedback of intonation, stress and rhythm". This is indeed the primary focus of the package: suprasegmentals. The decision to circumscribe the content area so narrowly is supported on the website of the package (http://www.betteraccent.com/papers/quotes_on_pronunciation.htm) with quotes from a number of experts in the field, such as Joan Morley, Marianne Celce-Muria, Joanne Kenworthy, even Alexander Graham Bell! This is very much in the spirit of the communicative language teaching approach, whereby communication is supposed to be maximized even at the expense of (phonetic) correctness. It is claimed that ill-pronounced individual segments will rarely hamper mutual understanding as much as incorrect prosodies: hence the emphasis on the latter.

The curriculum covered by the Tutor includes: *word stress, simple statements, wh-questions, general questions, repeated questions, alternative questions, tag questions, commands, exclamations, direct address, series of items, long phrases, tongue twisters*. The approach is heavily 'know-that'-oriented, with a lot of 'explanation' carried out in rather dense phonetic jargon; all of these features illustrated in Figure 7. ASR is used to display the learner's pitch (intonation) and intensity (loudness) graph alongside the model ones for visual inspection and comparison. No automatic evaluation is attempted.

With such a narrow focus and modest content, the package must compromise flexibility of application, of course. Notice also that the multimedia technology does not reach beyond audio play-back and input, with no graphical animation or video. Likewise, there is no phonetic transcription, articulatory diagrams, waveform comparison or traditional phonetic exercises (cloze, dictation, multiple choice, and the like). Thus, the package projects a rather austere image, as also transpires from the screenshot in Figure 7.

3.5. *ISLE*

Unlike all the CALL packages so far discussed, the *Interactive Spoken Language Education* (ISLE) was not a privately owned commercial venture, but rather a multinational (German-Italian-British) project, running between 1998 and 2000, heavily subsidized by EC funds, and coordinated by *Educational Concepts*, the R&D department of Ernst Klett Verlag. Three of the

partners were universities of Milan, Hamburg and Leeds, which again makes the project unique among those delivering pronunciation-oriented CALL software with a market potential.

According to the website, "The main objective of ISLE is to provide technical solutions to support training of spoken language communication. This will be achieved by developing computer based tools to support the training of speaking skills and by integrating such tools into existing multimedia-based language teaching software systems". The deliverable, however, whose demo can be freely downloaded from the Internet, is a stand-alone package apparently targeted at the busy manager in whom "the social climate of a classroom can easily produce psychological barriers to the training of elementary speaking skills in a foreign language". This can be seen even in the names of the entry screen components, such as "Travel Arrangements", "At the Airport", "In the Office", "At the Hotel" or "In the Restaurant". The content of the whole package, including dialogues, exercises and glossaries, is also adjusted accordingly. For example, we have a Paolo Rossetti arranging for his business trip to Manchester. The example screenshots all come from this module of the package.

After entering the program the learner must first calibrate the ASR engine by reading a story of the conquest of Mount Everest. This takes a few minutes. Choosing "Travel Arrangements" takes the user into the working area where he should first listen to a (video-less) dialogue between Paolo and a travel agency. The choice is between the captioned and sound-only modes. Then, a suite of tasks can be entered, conveniently divided into text- and pronunciation-based. Among the former, there are true-false ones, based on the dialogue, as well as translation, cloze, Q&A and correct-the-sentence. Interestingly, some of these are L1-sensitive: my demo version of the program happened to be one targeted at the Italian market; hence the prompt to "Thank you for ... with us" is 'volare'.

The "Oral Exercises" section offers read-and-repeat, listen-and-repeat, Q&A, Build the Sentence and Free Choice. The first of these is illustrated in Figure 8, the last but one – in Figure 9. What is of particular interest here is the 'improve' option: when the program decides that the learner mispronounced some sounds (no suprasegmental practice in this package), it offers advice and provides corrective practice, as shown in Figure 10, first kindly asking the learner "How strict should I be?". Upon testing, it turns out – somewhat expectedly – that with the 'strict' setting there is no way to make the ASR wizard satisfied. There is – again not unexpectedly – no guidance on how best to approach the recorded model; it is all the matter of hit-and-miss.

As can be seen from this short description of the functionalities of the package, as well as from the screenshots, despite claims to the effect that the program focuses on natural communication, the tasks are rather traditional, with repetition galore, comprehension questions and phonetic drills. While the learner can listen to a conversation conducted in a natural setting, he cannot himself engage in one in any form. And why should he ever need 'tanks' and 'ants' in the business context (see Figure 10)?

The ASR evaluation of sentences is characteristically unreliable, and the segmental ASR-assisted practice – tedious and unhelpful. The overall balance is in favour of procedural knowledge, with close to no explanation, no (phonetics) manual and no phonetic terminology. IPA transcription is used (sparingly) in the 'improve' menus. No technological gimmicks with waveform display

and adjustment are in sight. Generally, with no animation or video movement on the screen, the impression is that it is very traditionally rendered, despite the use of ASR. This, in turn, leads to the guess that the EC funding was not adequate to elaborate the content and function of the program any further.

3.6. *Tell Me More*

Finally, a CALL package where ASR technology has been used most effectively: to actually simulate a spoken dialogue between the learner and the computer. Auralog's *Tell Me More* is heavily advertised on the web as "the reference in foreign language learning, developing all linguistic skills: oral and written expression, comprehension, grammar and vocabulary". As far as pronunciation is concerned, it boasts "the exclusive S.E.T.S. technology (Spoken Error Tracking System) automatically detecting errors in pronunciation" as well as "3D phonetic animations" (see Figure 11). The program is sold in nine different language (L2) versions, including both British and American English, and three proficiency levels. There are also networked modalities of the software, with functionalities allowing teacher control and class management as well as student-teacher messaging and other tools. In what follows, however, I will describe the 'traditional' CD-ROM-based package.

As mentioned above, in *Tell Me More* the ASR technology is pushed to its current limits: (a) waveform display, (b) pitch tracking (see Figure 12), (c) learner input evaluation, (d) dialogue simulation. The latter proceeds by the program offering the learner a few printed options to read off the screen in response to the computer-initiated contextualized dialogue in an authentic setting, e.g. travel arrangements (Figure 13). The ASR engine tries to figure out which option was actually spoken, and reacts accordingly by responding to this user input. While this is far from an actual conversation, of course, the technique is reasonably robust and very motivating: the learner at last feels that what he says will change the following flow of communicative interchange. As mentioned above (2.6.), to achieve more along this path, ASR would have to feed into a functional AI component with L1 sensitivity and learner modeling. Such packages will not be available for... some time to come.

Tell Me More is definitely balanced towards procedural knowledge, with heavy emphasis on pronunciation (speech communication), although it is hard to make blanket statements for this package which appears on the market in so many different versions: proficiency-, L2-, learner-group-wise (there are dedicated business courses, for example). Even the title of the whole series changed over time, from *Talk to Me* in 1997; some older versions of the package are still available under this title. Unlike *Pronunciation Power* and *Connected Speech*, there is thus much less formal exposition of matters phonetic, no structural division of the program into phonetic fields such as segmentals, stress, intonation, and no phonetic terminology. In these respects *Tell Me More* resembles *ISLE*: speech communication in a naturalistic setting is at the centre of the package. Unlike in *ISLE*, however, the ASR engine does not attempt to identify specific errors in the learner speech input; rather the assessment is global. The acceptance level can be adjusted by the learner himself, with all the disadvantages described above (too strict or too lenient). The program is lavishly illustrated with good quality photos and videos (in newer

versions) and it shows all signs of professional graphics design. While the semi-transparent animation of the articulators (what is called "3D phonetic animations" on the web) may be little more than a gimmick at this stage of human-computer interaction, it does show us a glimpse of things to come in terms of educational applications of virtual reality (see Baldi below).

4. The future of pronunciation-oriented (EFL) CALL

Automatic Speech Recognition (ASR) is by far not the last word in human-computer interface design. As mentioned in the introduction, the pace of technological innovation in computer technology generally, and in natural language processing (NLP) in particular, is breathtaking. In this last part of the paper I can only try to briefly speculate about the impact on (EFL) CALL of some recent inventions and developments. Like before, the selection is of course heavily subjective, but – it is hoped – not quite irrelevant for the discussion above.

4.1. Text-to-Speech Synthesis

One area where the impact of technology on CALL is going to be felt soon is that of speech synthesis. Text-to-speech (TTS) synthesis by rule, whereby no previous human recording is necessary in any form, has reached human-like quality (cf. e.g. Dutoit 1997 and 1999). The high-end TTS engines are rather expensive now, and research to improve especially the prosodic properties of synthesized speech are still under way, but the technology is now reaching the stage where it can be applied to CALL, as the synthesized speech can actually function as a model of pronunciation, as well as in the now trite capacity of information deliverer[7].

The added bonus, compared to pre-recorded human speech, is that it is under stringent control of the designer in terms of practically all phonetic criteria: pitch (responsible for the impression of gender), tempo, intonation, timbre of voice, accent, loudness, etc. It would be technically rather easy to simulate a foreign accent, if need arose, for example to better demonstrate to the learner the areas which need improvement (e.g. final devoicing in Polish English). Keller & Zellner-Keller (2000a) note 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)".

Because TTS engines are tiny compared to audio recordings, the CD space recovered could be used for other multimedia components of CALL, such as video, for example (see below for video synthesis). And, naturally, speech synthesis is by far less expensive than recording a team of highly trained human speakers.

4.2. Face animation

One of the most successful applications of 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 ASR enhanced with the animated face ("Baldi", Figure 14) simulating phonetically realistic articulatory movements in real time. 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 build 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 15). 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. "Shrek" 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.

4.3. Multimodal man-machine communication

For truly multimodal human-machine interaction the machine would have to progress beyond simple (?) ASR – into the realm of automatic recognition of audiovisual speech. The AI TTSS ASR agent would be able to recognize and act upon (at least) the facial expressions of the computer user. This would not only aid communication generally through taking advantage of gaze, eye-brow movement, head positioning, and the like, but also – in the context of pronunciation teaching – make it possible to provide additional articulatory feedback to the learner concerning his lip position and movement in labial(ized) sounds, tongue-tip control in apico-dental fricatives or labio-dental contact in /f/ or /v/. Of course, to achieve this level of

video sensitivity highly sophisticated systems would have to be employed. Contemporary prototypes are nowhere near the required technological stage (see Figure 16 for a simple example). The area is full of vibrant research activity, however (see e.g. Granström, House & Karlsson 2002, or Scott 2001 for an accessible introduction), and the feeling is that we can expect significant breakthroughs quite soon.

4.4. Machine Translation (MT)

Machine translation might at first sight appear not to belong here, in a paper where pronunciation-oriented (EFL) CALL is discussed. It is indeed true that MT is seldom used in this context, even though one can envisage its creative applications in a grammar class, for example, where learners would try to induce the rules of the foreign language from the (usually risible) characteristic errors of an MT package. However, the impact of speech-enabled, or Speech-to-Speech (StS) MT, once it is perfected, can be enormous, not only for the business of pronunciation instruction, but for the whole world of foreign language teaching (FLT) and learning. In the words of Crystal (2001:227): "We can also envisage the translating telephone, where we speak into a phone, and the software carries out the required speech recognition, translation, and speech synthesis, enabling the listeners to hear our speech in their own language [...] Such a world is, of course, a very long way off". It is enough to have a look at Ectaco's web pages (http://www.ectaco.com/) to appreciate that the envisaged world has already arrived: while the Russian company's translator is still rather primitive (in terms of device size, range of languages translated, vocabulary size, noise robustness, etc.) it does translate speech to speech in real time with quality quite adequate for a tourist or businessman ordering a meal in a restaurant or air tickets in a travel agency. Thus, I believe, Crystal's "way-off" nightmare is more immediately threatening than he ever thought: "in a world where it is possible to translate automatically from any one language into any other, we have to face up to the issue of whether people will be bothered to learn foreign language at all" (ibidem; see also Cribb 2000 for similar conclusions).

5. Conclusions

With the currently fashionable 'focus on form' in foreign language teaching (EFL in particular; see, for example, Doughty & Williams 1998 or Ellis 2001) the role of pronunciation-oriented CALL software is bound to grow in the process of phonetics teaching and learning, both in the classroom context and outside it. Teachers will delegate onto computers some of the more tedious tasks involved, such as drilling as one technique of skill-getting. It seems that the unconstrained smoothly-flowing spoken foreign-language dialogue with the computer will not become reality for some time to come. So, until computer AI improves significantly, truly communicative activities will not be used as a vehicle for practicing pronunciation. But at least there is a chance that the employment of (EFL-aware) speech recognition, text-to-speech synthesis and certain elements of artificial intelligence will gradually transform the boring

phonetic 'drill-and-kill' procedure into an exciting, multimedia, interactive 'drill-and-thrill' adventure.

The audio channel of communication between human and machine, which is now opening, both in FLT and outside it, is an additional boost for the growing anthropomorphization of the computer. After all, with which other creatures, natural or artificial, can we communicate by voice? The computer will unavoidably grow its own persona around it. The FLT learner will take it more and more for granted that he can intelligently communicate with this persona in the foreign language. He will react to it more and more on the affective level, as well as intellectually. He will like it, or hate it, as the case may be. He will look to it for help, advice, praise and criticism. He will count on its inherent intelligence and wisdom. What impact this attitude will have on foreign language learning and teaching remains to be discovered. My guess is that it will be enormous.

Notes

1. This text is based on my lecture to the SCE Foreign Language College (http://www.nkjo.szczecin.pl/) in Szczecin on January 16th 2003. While the overall organization reflects that of the lecture, the text is of course not a mirror image of the latter, if only because it cannot contain the rich multimedia content presented in the College. The original text was written in April 2003. Due to adverse circumstances its publication in Szczecin has been suspened. I believe, however, that the main theses of this paper remain in force. Links to respective web pages were checked and updated 15 December 2004. Otherwise, with very minor editorial changes, the text appears in its original form. I am grateful to Dr. Jarek Krajka for offering TEwT as the venue for its publication.

2. In this context a multitude of books (e.g. Jenkins 2000), articles (e.g. Sobkowiak 2003) and conferences (e.g. LM34's workshop on LFC: http://elex.amu.edu.pl/ifa/plm/2003/index.htm) might be invoked.

3. Somewhat paradoxically, as early as at the beginning of 1980s Sinclair's ZX Spectrum could (with pains) flash user-designed IPA on the TV screen, which functioned as its VDU (Video Display Unit).

4. This may not be true of some speech assessment and practice software mostly used in the (native-language) clinical setting. This is, however, usually rather narrowly targeted at, say vowel quality, whereby the user is trying to match vowel formant positions in two-dimensional diagrams with the model ones.

5. The shape of TH is somewhat nonstandard, though: /đ/.

6. Even if "experiments have shown that a visual display of the talker improves not only word identification accuracy [...], but also speech rhythm and timing" (Ehsani & Knodt 1998:52).

7. See *Filoglossia*, a CALL package with Greek as a foreign language, which already employs TTS synthesis: http://www.ilsp.gr/filoglossia_plus_eng.html, or WordPilot from http://www.compulang.com, which also has this feature.

Bibliography

Aist, G. 1999. "Speech recognition in computer-assisted language learning". In K.Cameron (ed.). 1999. 165-81.

Asher, R.E. & J.M.Y.Simpson (eds). 1994. *The encyclopedia of language and linguistics*. Edinburgh: Pergamon Press.

Bernstein, J. & H. Franco. 1995. "Speech recognition by computer". In N.Lass (ed.). 1995. 408-34.

Cameron, K. (ed.). 1999. *CALL: media, design and applications*. Amsterdam: Swets & Zeitlinger.

Carey, M. 1998. "But is it interactive and does it work? A review of some CALL English pronunciation CD-ROMs". In In T.Ottmann & I.Tomek (eds).1998.

Chen, Hao-Jan H. 2001. "Evaluating five speech recognition programs for ESL learners". Paper presented at ITMELT'2001, Hong-Kong, 9 Nov 1999.

Cole, R. & V. Zue. 1996. "Spoken Language Input". In: R. Cole et al. (eds). 1996. *Survey of the state of the art in human language technology*. [http://cslu.cse.ogi.edu/HLTsurvey/ch1node2.html#Chapter1; last accessed 15.12.2004]

Cribb, V.M. 2000. "Machine translation: the alternative for the 21st century?". *TESOL Quarterly 34.3*, 560-9.

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

Deroo, O. *A short introduction to speech recognition*. [http://www.babeltech.com/download/SpeechRecoIntro.pdf; last accessed 15.12.2004].

Derwing, T.M., M.J. Munro & M. Carbonaro. 2000. "Does popular speech recognition software work with ESL speech?" *TESOL Quarterly 34.3*, 592-603.

Doughty, C. & J. Williams. 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; last accessed 15.12.2004**]**

Ehsani, F. & E. Knodt. 1998. "Speech technology in computer-aided language learning: strengths and limitations of a new CALL paradigm". *Language Learning & Technology 2.1*, 45-60.

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

Eskenazi, M. 1999. "Using automatic speech processing for foreign language pronunciation tutoring: some issues and a prototype". *Language Learning & Technology 2.2*, 62-76.

Fotos, S. (ed.). 1996. *Multimedia language teaching*. Tokyo: Logos International.

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

Hiller, S.M. & E.J. Rooney. 1994. "Teaching pronunciation with computer aids". In R.E. Asher & J.M.Y. Simpson (eds). 1994. 4528-35.

Hunyadi, L. et al. (eds). 1998. *ALLC/ACH'98 conference abstracts*. Debrecen: Lajos Kossuth University.

Jenkins, J. 2000. *The phonology of English as an international language*. Oxford: Oxford University Press.

Kaliski, T. 1992. "Computer-assisted language learning (CALL)". In P.Roach (ed.). 1992. 97-109.

Kaltenboek, G. 2001. "A multimedia approach to suprasegmentals: using a CD-ROM for English intonation teaching". In J.A. Maidment & E. Estebas-Vilaplana (eds). 2001. 19-22.

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

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

Kern, R. & M. Warschauer. 2000. "Theory and practice of network-based language teaching". In M. Warschauer & R. Kern (eds). 2000. 1-19.

Lambacher, S.G. 1996. "Spectrograph analysis as a tool in developing L2 pronunciation skills". In M. Vaughan-Rees (ed.). 1996. 32-35.

Lass, N. (ed.). 1995. *Principles of experimental phonetics*. New York: Mosby.

Lecumberri, G., M. Cooke & J.A. Maidment. 2001. "Automatic feedback on phonemic transcription". In J.A. Maidment & E. Estebas-Vilaplana (eds). 2001. 15-18.

Maidment, J.A. & E. Estebas-Vilaplana (eds). 2001. *Proceedings of the Phonetics Teaching and Learning Conference*, April 5-7 2001, London. London: UCL Dept of Phonetics and Linguistics.

Nagy, T., P. Furkó & A. Tóth. 1998. "Computers and multimedia applications in teaching pronunciation". In L.Hunyadi et al. (eds). 1998.187.

Ottmann, T. & I. Tomek (eds). 1998. *Proceedings of ED-MEDIA/ED-TELECOM 98*. Chartlottesville,VA: AACE.

Pellegrino, F., C. Fressard & G. Puech. 2001. "Teaching phonetics with a multimodal Internet site". In J.A. Maidment & E. Estebas-Vilaplana (eds). 2001. 31-34.

Pennington, M.C. 1999. "Computer-aided pronunciation pedagogy: promise, limitations, directions". *CALL 12.5*. 427-40.

Roach, P. (ed.). 1992. *Computing in linguistics and phonetics*. London: Academic Press.

Rudnik, O. 1998. *ASR in EFL CALL*. Poznan: Unpublished IFA MA thesis.

Scott, N. 2001. "Intelligent user interfaces". *Information Impacts Magazine, March 2001*. [http://www.smartnet.co.nz/res-articles/intelligent-interfaces.htm; last accessed 15.12.2004]

Sobkowiak, W. 2003. "Why not LFC?". In W. Sobkowiak & E. Waniek-Klimczak (eds). 2003. *Dydaktyka fonetyki języka obcego*. Zeszyty Naukowe Państwowej Wyższej Szkoły Zawodowej w Koninie. Proceedings of the Wąsosze Conference (http://www.pwsz.konin.edu.pl/ins-neo.lm?tresc=ins-neo/instytut_neofilologii_wasosze2.html) on teaching foreign pronunciation, 10-12.5.2002. [English translation here: http://elex.amu.edu.pl/~sobkow/wasosob4.doc]

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

Vaughan-Rees, M. (ed.). 1996. *Speak out: newsletter of the IATEFL Pronunciation Special Interest Group (PronSIG)*. London: IATEFL.

Warschauer, M. 1996. "Computer-assisted language learning: an introduction". In S.Fotos (ed.). 1996. 3-20.

Warschauer, M. & D. Healey. 1998. "Computers and language learning: an overview". *Language Teaching 31*. 57-71.

Warschauer, M. & R. Kern (eds). 2000. *Network-based language teaching. Concepts and practice*. Cambridge: CUP.