M.-J. Hamel, and Handley, Z. (2002). On the use of speech synthesis in CALL. In Instill-CALICO*, Davis (USA), March 26-30*

On the Use of Speech Synthesis in CALL (provisional version)

M.-J. Hamel & Z. Handley
CCL, UMIST
Manchester, UK

1 What is speech synthesis?

In very simple terms, speech synthesis is the process of making the computer talk. There are several different classes of speech synthesis. The major distinction between them is the type of input that they support. The major classes of speech synthesis are:

- Concept-To-Speech (CTS) (message-to-speech)
- Text-To-Speech (TTS)
- Re-Synthesis (analysis/synthesis)

Concept-to-speech:
Takes concepts as its input and attempts to not only speak but also generate the phrases to utter.

Text-to-speech:
Takes raw text as input and aims to mimic the human process of reading. (D'Alessandro, 2001: example 36)

Re-synthesis:
Takes naturally produced utterances and modifies them through the application of digital signal processing techniques. May also take as input the output of speech synthesisers.

2 A technology little heard of in CALL...

Despite the fact that more than a decade ago some like Sherwood 1981, Stratil et al. 1987, Esling 1992, Last 1989, already saw several advantages in the use of speech synthesis in language learning and teaching, such technology until recently remained little heard of. Speech technologies began to be fully exploited in CALL applications during the mid-1990s. However, it has been observed that, from an NLP point of view, such applications mostly remained silent (Hamel 1998). This statement is even truer of speech synthesis which is still hard to find today integrated in CALL applications, whether commercial products or research prototypes. In this paper, we do not wish to concentrate on the reasons why speech technology has been ignored so far but rather focus our discussion on its advantages and possible/actual uses in CALL.

3 Speech synthesis in CALL: why?

As said, speech synthesis is initially perceived as a tool with potential beneficial uses in language learning and teaching. Last 1989, for instance, describes speech synthesis as "A technology which is becoming available which may overcome both the problem of time during which the learner is waiting for his response to come on-line, and also the considerable time input by the teacher in preparing the audio and computer-file based material for rests of this kind" (Last, 1989: 144). About ten years later, speech technology is said to be mature enough (systems are robust and their output reliable: Hamel 1998) to be exploited for teaching and learning purposes in CALL applications. Skrelin and Volskaja

(1998), for instance, highlight the fact that “Nowadays we have at our disposal the technical means for accurate text analysis” (Skrelin and Volskaja 1998: 21).

More recently, the advantages and the possible uses of speech synthesis in CALL have attracted the attention of technology specialists themselves. Dutoit (1997), for instance, suggests that high-quality TTS synthesis could be used in language learning, although he goes on to say “To our knowledge, this has not been done yet, given the relatively poor quality available with commercial systems, as opposed to the critical requirements of such tasks” Dutoit (1997: 31). Keller and Zellner-Keller (2000) give more specific reasons why they believe that speech synthesis should be used in language learning, which we shall present in the following paragraphs.

Quite early, Sherwood (1981) already saw certain teacher-oriented advantages in the use of speech synthesis (over the use of audio-tapes) in language learning and teaching, advantages which can also apply to CALL. According to the author, it is easier to:

- type text than record voice;
- edit/change text than re-record voice;
- navigate through textual samples within a database than recorded samples on a tape;
- generate new (feedback) examples.

Similar teacher/developer-oriented advantages are later expressed by Skrelin and Volskaya (1998) with relation to CALL. According to the authors, the use of speech synthesis in language-learning systems would:

- free up memory for other media;
- free up screen space for other media.

Raskind and Higgins (1995) saw several learner-oriented advantages:

- Generally such assistive technologies increase learner enthusiasm, interest and motivation.
- With respect to error identification, listening to a text has been shown to be better than reading a text.
- Evidence from experiments in the use of talking word processors – made more revisions and produced longer texts.
- May help improve reading, spelling and writing

Finally, Keller and Zellner-Keller (2000) saw the following both learner and teacher-oriented advantages in the use of speech synthesis in CALL. Indeed for the authors, the speech synthesiser:

- is an “indefatigable substitute native speaker” (op. cit.: 111);

- it is not human and therefore perceived as being non-judgemental (This is particularly useful with illiterates, as illiteracy is often stigmatised.)

- can be used to produce special types of examples which are useful in language teaching. Indeed, “Speech synthesis allows repetition at will, as well as the presentation of exercises specifically adapted to the needs of the student, plus the

> creation of sound examples that could not be produced by a human being (e.g., speech with intonation, but no rhythm)." (op.cit.: 110)

4 Speech synthesis in CALL: what for?

Having mentioned a few of the advantages related to the use of speech synthesis in CALL, we will now look at some possible and actual uses of such a technology in CALL applications.

CALL applications have emerged from the general need in language learning for "self-paced interactive learning environments" which provide "controlled interactive speaking practice outside the classroom" (Ehsani & Knodt, 1998).

Within this perspective, speech synthesis can contribute to CALL by providing electronic systems with 'a voice', a voice which can be used for several purposes and in several contexts. Here are few of these.

4.1 Talking dictionaries

This is the most popular CALL/use for CALL purposes application exploiting speech synthesis. A lot of work has focused on the development of talking dictionaries probably because it is (one of) the easiest applications to develop. The automatic access to spoken output provides the learner with an instant pronunciation model that can be imitated - imitation is a good form of lexical reinforcement (Myers 2000). Talking dictionaries are also used to develop a more conscious awareness of the relationship between both the graphic and the phonic form of lexical items (sometimes the IPA forms too), a relationship which is not always straightforward in some particular languages and/or for some learners. Examples of such talking dictionaries can be found in various languages, namely French (Hamel, Nkwenti-Azeh and Zahner, 1996, Hamel 1998), Russian (Skrelin and Volskaya, 1998), Kurdish (Fatah, Elturan and Durroie, 1998) and Breton (Mercier et al, 2000; Siroux et al, 1998).

The Breton (ibid) talking dictionary contains 35 000 entries with definitions. The user can navigate through both the French and Breton lexical forms. IPA transcriptions and text-to-speech output are provided for all dialectal variations of Breton pronunciation. The system does not use full text-to-speech processing as it takes the IPA transcription as input The dictionary can be searched using regular expressions in a grep-like style.

4.2 Talking texts (On-line reading)

The is the second most obvious use of speech synthesis in CALL applications where words, sentences, portions of texts or entire texts can be selected and instantly read aloud to the learner in order to support his/her reading comprehension activities. It has been suggested that providing speech output for on-line reading will improve the navigation of CALL programs, because many are in hypertext environments which are full of colours and links which distract from the material and impede/slow down navigation (Moisa and Ontanu, 1999). Examples of CALL applications exploiting speech synthesis for talking texts are described in Hamel (1998) and Godwin-Jones (2000). A current E.C. funded CALL project, FreeText[1] [footnote: project nu. + partners + http], integrates the text-to-speech synthesiser,

[1] FreeText (French in Context: An advanced hypermedia CALL system featuring NLP tools for a smart treatment of authentic documents and free production exercises): EU project IST-1999-13093 involving Centre for Computational Linguistics, UMIST, Manchester, UK, Département de Linguistique,

FIPSvox (Gaudinat and Wehrli, 1997), for talking text purposes as described in Hamel (1998) for the SAFRAN project (Hamel 1998, Hamel and Vandeventer 2000, Vandeventer and Hamel 2001).

4.3 Dictation

Dictation is a language learning writing activity focusing on the learner's spelling skills. As mentioned in the above section about talking dictionaries, its overall aimis to reinforce the phoneme-to-grapheme relationship. The use of speech synthesis for dictation activities in CALL systems allows the teacher to input the texts very easily and the learner to listen them at his/her own pace, in a less/non stressful learning environment (IRISA, 2000; Mercier et al, 2000). Such dictation systems often also integrate an error diagnosis device based on simple pattern-matching techniques or NLP techniques for automatic error detection. SAFexo (Hamel 1998), DICTOR (Santiago-Oriola, 1999) and ORDICTEE (Mercier et al. 2000) are examples of CALL applications which integrate or are entirely dedicated to such dictation activities.

With ORDICTEE (IRISA, 2000; Mercier et al, 2000), for instance, the computer takes on the role of the teacher and dictates the text to the learner whilst the learner is typing at the keyboard. A particular feature of this system is that it adapts the pace to learner; the speed of dictation adapts itself to the learner’s typing speed. Once the dictation has been completed, the computer system also corrects the text and displays the errors to the student, however no constructive feedback on how to correct the errors is given. The learner can manipulate the following parameters of the synthesised speech: “the speed of elocution, the voice pitch and volume” (Mercier et al, 2000: 147).

4.4 Pronunciation training

Pronunciation training is particular in that it typically involves a close and quite repetitive one-to-one relationship between the learner and his/her teacher. It is an activity that is often inhibiting for the learner in a classroom situation and demands individualised attention from the teacher with all participants. CALL in that sense, because of its individual, self-paced and indefatigable nature, is seen as a well-suited candidate for this kind of activity provided that it is speech-enabled. A lot of speech-enabled CALL systems developed so far focus on speech recognition. Here, we demonstrate that speech synthesis can (and has) also been put to good use in the teaching of pronunciation.

Generally speaking, the use of speech synthesis to support pronunciation training allows the learner more control over the output presented to him/her as he/she can typically "slow down stretches of spoken language at will, which eases familiarisation and articulatory training with novel sound sequences" (Keller and Zellner-Keller, 2000: 110). He/she can also repeat and imitate utterances over and over again as the synthesiser will not tire like a human would, since the tool is, as said earlier on, an “indefatigable substitute native speaker” (Keller and Zellner-Keller, 2000: 111). It has also been said that learners prefer judging their degree of
phonetic accuracy by repeatedly listening to the models and their own
auto-evaluation (Reeser 2002).

More specifically, speech synthesis can be exploited for pronunciation training at segmental (practice of individual and combined phonemes) and supra-segmental (practice of intonation and prosody) level. It can also be used to reinforce, as said in the previous sections, the relationship between the phonic and the graphic form. We will see reasons why in more details here.

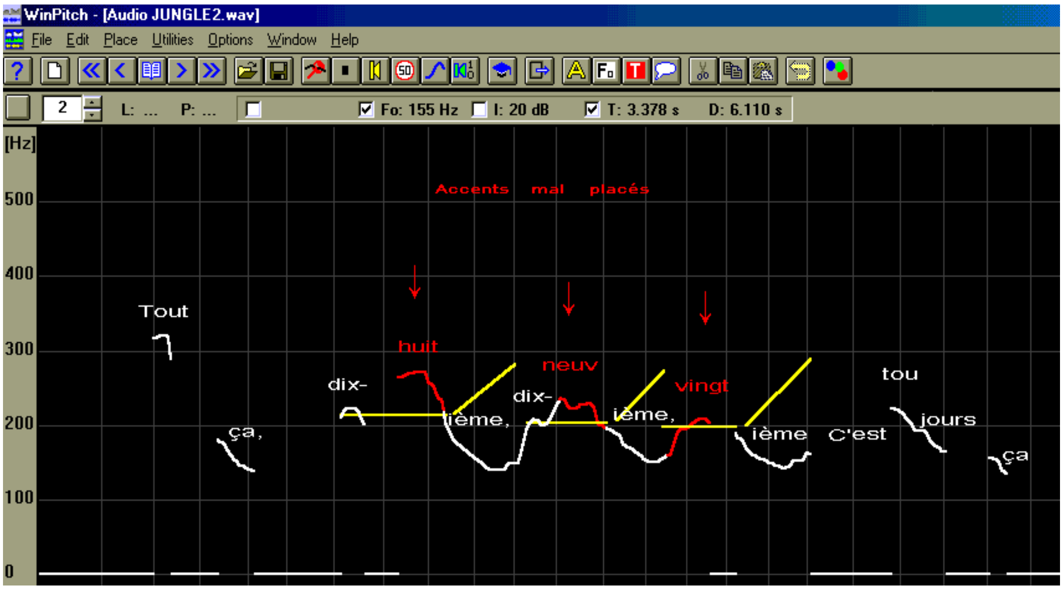
4.4.1 Practice of intonation and prosody

According to Skrelin and Volskaya (1998) "the advantage of the synthesised intonation contours compared to those realized by the speaker is that their melodic pattern is more formalized. Thus the generalized contours free from additional emotional colouring are used as models." (Skrelin and Volskaya, 1998: 24). Similarly, Knoerr (2000) quotes the work of Feldman (1977) who experimented with using speech synthesis techniques to produce examples of intonation for use with learners. The conclusion of this work was that learners found it easier to discriminate between the simplified examples produced through synthesis. Knoerr (2000) then reports on the use of the visualisation software Videovoice to teach intonation. The software which allows the presentation of pitch contours was used to teach learners intonation. In particular, she reports on the use of this software in repetition exercises to teach learners the 10 basic (stylised) intonation patterns of French. The results of the study showed that in all cases the students improved their intonation and in some cases they improved it considerably. It also showed that in most cases the students spent more time than usual practicing outside the classroom.

Bonneau et al (2000) suggest that waveform manipulation can be used in conjunction with a text-to-speech system to help learners perceive intonation better. In particular, they describe how the signal analysis tool, WinSnoori could be used in conjunction with the Klatt synthesiser. Here, to focus the learner's attention on the prosody of utterances, the teacher can manipulate the fundamental frequency (F0) of the utterances produced by the speech synthesiser. A global multiplicative factor may be applied to the whole utterance or selectively to sections of the utterance in order to focus the learner's attention on the fall or the rise of the intonation. In order to improve the learner's perception of stress, the teacher can manipulate stress patterns, for example he/she can invert them. Like Nagano and Ozawa (1990), the authors also suggest that "The transformations can be realised on the learner's voice, allowing him/her to better apprehend the target pronunciation" (Bonneau et al, 2000: 77).

Waveform manipulation to hear own voice:

This technique is based on the idea that listening to another persons voice distracts attention from the features specific to intonation (Yoram and Hirose, 1996). It was originally proposed by Nagano and Ozawa (1990) for teaching English to Japanese learners. Utterances which have been recognised and found to deviate from the pronunciation model are analysed, then manipulated and then re-synthesised with the correct pitch contour (for example) so that the learner can hear themselves pronounce the utterance correctly. The results of Nagano and Ozawa's (1990) experiments showed that manipulating the learner's own utterances was indeed more effective than using native speaker's voices as feedback (Aist,1999: 170). The method has hence subsequently been adopted by Yoram and Hirose (1996), Germain and Martin (2000) and Germain-Rutherford (2001). In particular, Germain and Martin (2000) and Germain-Rutherford (2001) use the speech analysis software WinPitch LTL (a version of WinPitch that has been adapted for language teaching) to achieve it. These types of waveform manipulations obviously involve the use of speech recognition techniques.



WinPitch - [Audio JUNGLE2.wav]

☑ Fo: 155 Hz □ I: 20 dB ☑ T: 3.378 s D: 6.110 s

Figure X Feedback through speech synthesis taken from Germain-Rutherford (2000)

4.4.2 Practice of individual and combined phonemes

Speech synthesis can be exploited to practice the pronunciation of individual or combined phonemes. The synthesiser will typically be used to present all individual/combined sound segments/stimuli to the learner, segments/stimuli which would have been stored in a database in textual format. SAFRAN's pronunciation tutor, SAFexo (Hamel 1998), focuses on this kind of practice, a practice where essentially, three main types of activities are involved: auditory discrimination,
simple repetitions and repetitions involving phoneme/segment manipulations. In all three cases, the speech synthesiser is used as a model to imitate, and asa model to compare their own pronunciation/results with.

Waveform manipulation techniques can also be used to teach the phonetics of the language at the segmental level. Bonneau et al (2000) also describe the use of waveform manipulation techniques to teach segmental aspects of language/to enable learners to perceive segmental aspects of language better. Two techniques are described: spectral enhancement and slowing speech rate down. Spectral enhancement is used to make stops, bursts and fricatives easier to perceive. In addition, in order to make the speech easier to perceive it may be slowed down globally or selectively (i.e. only those sounds/speech sounds/segments which are difficult to perceive).

A modification/extension of this approach is described in Protopapas and Calhoun (2000). They describe how the approach can be used to teach auditory discrimination. To improve discrimination they suggest that waveform manipulation can be used to exaggerate the distinction between two phonemes which are wrongly perceived as being the same (for example, Japanese learners of English have difficulty in making the l/r distinction) until learners can perceive the difference. Learners are then gradually presented with utterances which are nearer and nearer to natural productions.

4.4.3 Relationship between phonic and graphic forms

From a language to another, as said earlier, the relationship between the phonic and the graphic form is not always straightforward as the correspondence between both sounds and letters can be quite indirect: one sound can corresponding to more than a letter, a sound can be mute in some contexts (assimilation, epenthesis) or appear in some others (liaison), etc. The learner who has to adapt to a new writing system in addition to a new phonetic one, will need to focus on the relationship between these two systems even more. Dictation activities are one way of reinforcing this aspect. Other possible ways of doing this are via games such as 'talking cross-words', 'talking-hang-man', talking 'fill-in-the-blanks'. Such activities are currently part of the pronunciation tutor SAFexo (Hamel 1998). While attention can be given to the graphic form, it is also possible to think of activities for which attention is given to the phonic form. Again, ways of doing this are via games such as 'anagrams', 'tongue-twisters' and karaoke' where the learner is presented with particularly challenging written sentences which he/she has to read aloud and compare with the output of the speech synthesiser (unfortunately the 'karaoke' does not yet sing!). Such activities are currently part of the pronunciation tutor SAFexo (Hamel 1998).

4.5 Dialogue partner

In opposition to the uses described above, Egan and LaRocca (2000) argue that “Synthetic speech has a role in language learning as much as ASR [Automatic Speech Recognition], but not as a model to imitate but as a means to complete tasks with full spoken interactions” (Egan and LaRocca, 2000: 5). In their opinion, the output of speech synthesisers is not of good enough quality to allow learners to imitate it. They do however believe that speech synthesis has a place/a role to play in CALL. They believe that the low storage requirements and generative power of text-to-speech synthesis should be harnessed in order to provide dialogues/conversational practice; responses in dialogues are unpredictable and maybe infinite in number, it is therefore both difficult to predict and to store all possible responses in the form of digitally recorded human speech to learner utterances.

5 Visually enhanced speech (synthesis)

Visual output can also be used in conjunction with speech synthesis. The idea is the it enhances the auditory stimulus and therefore might provide higher chances of success as far as listening comprehension is concerned which itself might lead to higher chances of success as far as pronunciation is concerned.

In general, until now these techniques have been restricted to use with people with speaking disorders (SpeechAssist, VideoVoice) or have only been used in CALL applications in conjuction with speech recognition (Auralog, Dyned products) without the use of TTS.

Typically, two main types of visual reinforcement can be used in conjunction with speech synthesis in CALL: acoustic display and articulatory display.

5.1 Acoustic display

Sounds can be displayed acoustically in three different ways: waveform, spectrogram and pitch contour display.

A waveform display is an electronic visual representation of the amplitude (the intensity) of sounds. Such visual output is found in commercially available products such as Dyned and Auralog series which do not make use of speech synthesis.

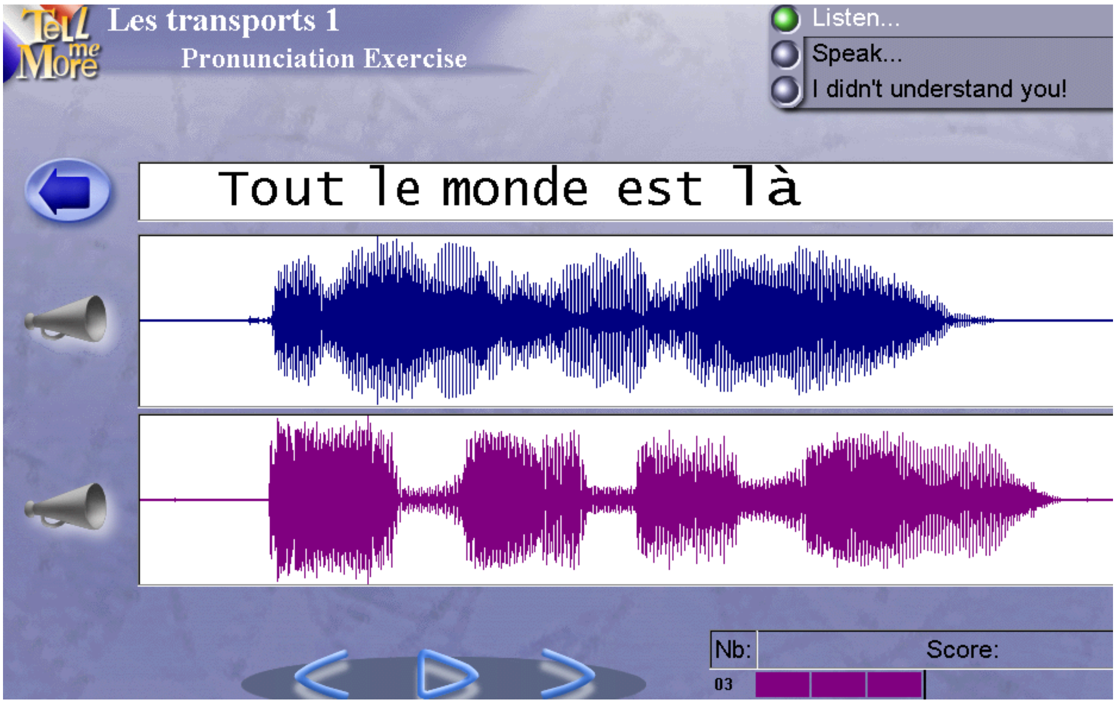


Fig. 5.1 Waveform extracted from Auralog product

Reeser (2002: 423) makes the following criticism of these displays: "pronunciation models are very impressive looking, but it was not always clear … how the visual support can help them [students] improve their pronunciation."

A spectrogram (sonogram) display is an electronic visual representation of the formants and transitions traces left by voiced speech and noise.

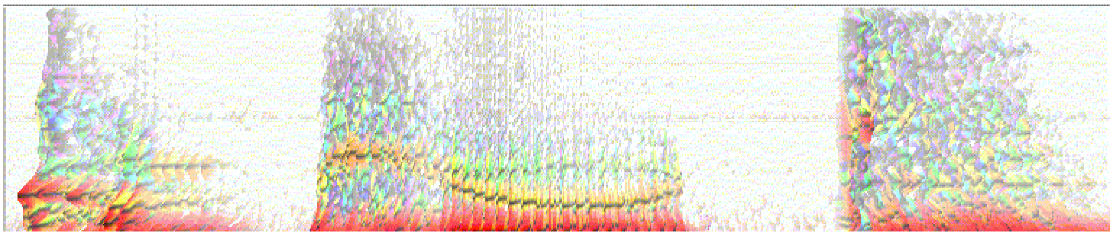
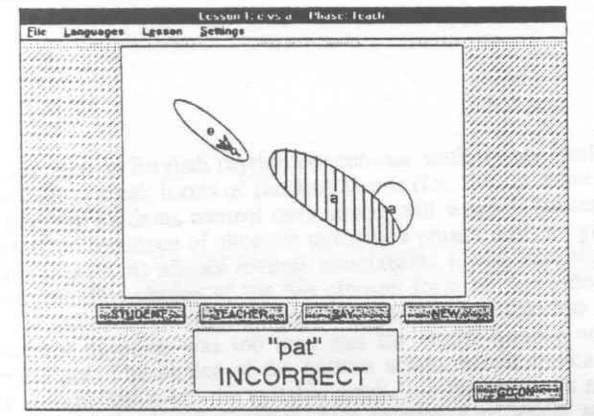


Figure X A spectrogram

Spectograms have been interpreted, to show, as in this example, a vowel target as in SPELL project.



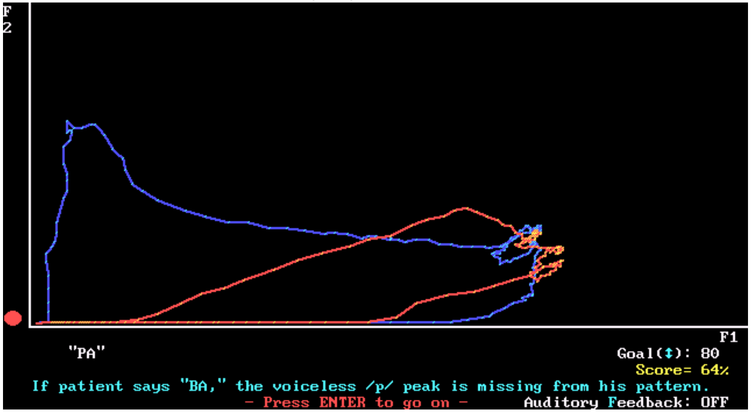
"pat"
INCORRECT

Figure 3. An example of the user interface for the vowel teaching module for a student learning English.

Figure X Vowel target representation in the SPELL project (Hiller et al,1994: 56)

VideoVoice also interprets the output of spectrograms in a similar way.

Formant (vowel/word) display in VideoVoice



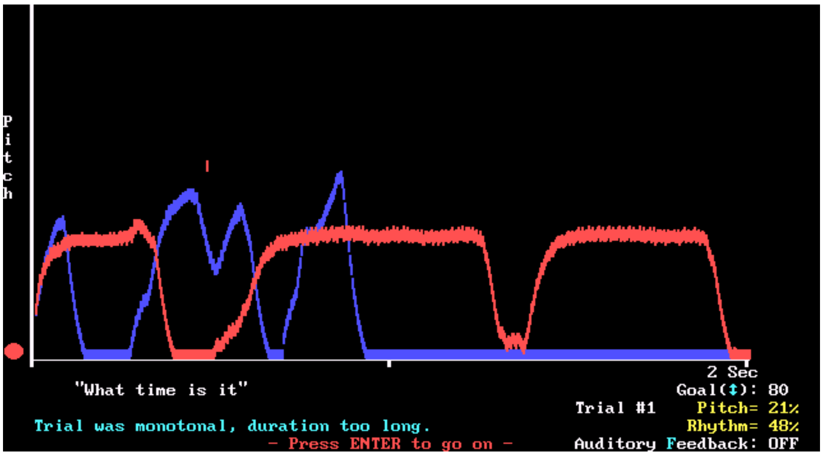
Goal(↕): 80
Score= 64%

If patient says "BA," the voiceless /p/ peak is missing from his pattern.

- Press ENTER to go on -

Auditory Feedback: OFF

A pitch contour display is an electronic representation, in form of a graph, (as a intonation/melodic curve) of the fundamental frequency (F0) of sound/speech against time.



Goal(↕): 80
Trial #1
Pitch= 21%
Rhythm= 48%

Trial was monotonal, duration too long.

- Press ENTER to go on -

Auditory Feedback: OFF

Fig. X Pitch model vs learner production in VideoVoice

Pitch contours have been interpreted and displayed a tunnels in the SPELL project:

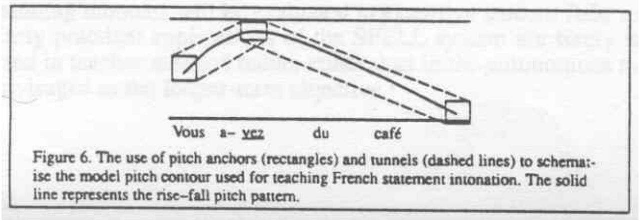


Figure 6. The use of pitch anchors (rectangles) and tunnels (dashed lines) to schema:-ise the model pitch contour used for teaching French statement intonation. The solid line represents the rise–fall pitch pattern.

Figure X Pitch tunnels in the SPELL project (Hiller et al, 1994: 60)

Komissarchick and Komissarchick (2000: 86) point out the fact that “There exists a trade-off between how rudimentary the analysis of a user’s speech and how easy it is for an average user to understand the resulting visual feedback. For example, speech spectrogram is easy to calculate, but “Teaching students and teachers what these [spectrographic] displays mean might take longer than the pedagogical potential their use might warrant.”

Hence, these visualisation techniques are limited by the learner’s ability to interpret the different displays and to distinguish an acceptable deviation from the model pronunciation from an unacceptable deviation. To overcome these limitations, these visualisation techniques are often coupled with speech recognition techniques used to identify whether the correct segments had been uttered and to what extent these segments deviated from the models in terms of segmental and supra-segmental aspects. But we believe that speech recognition is less reliable than speech synthesis and that the latter can be used as/more effectively if combined with “displays that are useful, easy to interpret, and that assist in language learning.” (Price, 1999)”.

5.2 Articulatory display

Articulatory features which referred to as 'visemes'. A viseme is a representation of the positions of the articulatory organs during the pronunciation of a phoneme. Visemes are therefore by definition more 'visual'. Therefore the idea of showing the learner articulatoty features instead of acoustic ones seems to agree better with Price's idea. Indeed, visual speech offers the learner with these additional advantages:

- audio-visual speech is more robust (Massaro and Cole, 2000: 157) [because relies on two mode of representation/feedback where one can be used to desambiguise the other];
- audio and visual speech provide complementary information (Massaro and Cole, 2000: 157) [as it allows learners to see how words are articulated not just hear “to pedagogically demonstrate correct articulation” (Massaro and Cole, 2000: 159)];
- “indefatigable substitute native speaker” you can see;
- adds realism to language learning.

We can distinguish two types of articulatory display or visible speech: 2D and 3D heads.

5.2.1 2D talking heads

2D talking heads typically show a simple, skin-less cross-section/sagittal section of the articulatory speech. Such 2D talking heads provide significant help/information/feedback to learners because in general most articulatory behaviour can be shown in this view/most articulatory distinctions are made along this plane.

An example of a 2D head is FREDA developed at UMIST (Mumford 1998) for the SAFRAN project (Hamel 1998). This 2D head is used in conjunction with the TTS synthesiser FIPSvox (Gaudinat and Wehrli, 1997). A set of 30 visemes is associated with the set of 36 French phonemes and the audio-visual output is displayed in real-time.

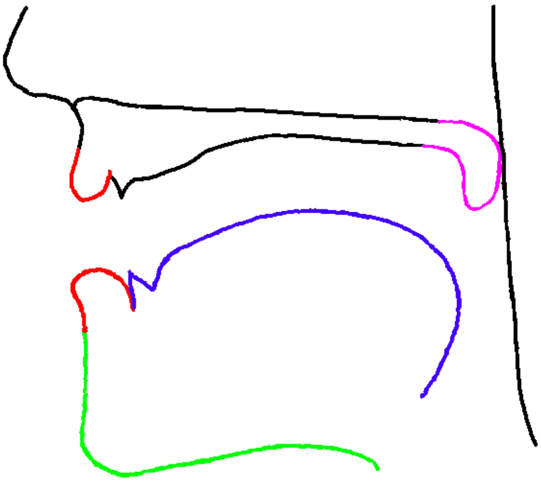


Figure 5.2.1 [p]/[b] viseme in FREDA (Mumford 98)

5.2.2 3D talking heads

Innovative functionalities make talking heads particularly useful in language learning. These functionalities are namely that:

- the head's is see-through so learners can see articlulatory behaviour that would otherwise be hidden behind the lips/cheeks;
- like audio speech such visual speech can be exaggerated to focus attention on the important/most salient characteristics of a given segment.

As both of these functionalities can be shared by both 2D and 3D head, the following characterises 3D heads:

- the head can be pivoted/turned around so that the learner views the head from the back – this means that the learner see the tongue and other articulators moving in the direction that he/she perceives his/her own to be moving.

Indeed, 3D talking heads can be rotated in all directions hence providing a more comprehensive view of the articulatory speech than 2D heads.

“Currently there are at least English (UCSC, Miralab, HMS, VisLab, DEC, BT, an so on and so forth), Swedish (KTH), Japanese (sony and NTT) and French (ICP and INA) speaking audio-visual speech synthesizers, so called talking heads” (Paula Web Survey).

Figure 5.2.2 3D head by Le Goff et al (1994)

6 The type of speech synthesis being used in CALL

In all cases where it is possible to establish the source of the speech synthesis system, these have been borrowed from other applications or are systems developed for general purpose. Here are some examples found in the literature:

- Mercier et al (2000) Breton: most likely that this system was borrowed from CNET (FT R&D);
- Keller and Zellner-K (2000) refers to the different versions of the Klatt synthesiser inc. DEC-Talk;
- Bonneau et al (2000) re-use of Klatt synthesiser and WinSnoori for editing and re-synthesis;
- Osborne, M (2000) describes how you can reuse any speech technology TTS and ASR that is compliant with Microsoft Speech API;
- Skrelin and Volskaya (1998) they are re-using the Russian TTS system developed by (Bondarko, L V, Skrelin, P A, and Volskaya, N B et al (1996) = RUSVOX;
- Santiago-Oriola (1998) experiments with TELEVOX;
- Hamel (1998)'s SAFexo uses FIPSVOX;

Yoram and Hirose (1996) use the Klatt synthesiser.

The same is true of the signal processing/speech analysis/synthesis tools. Both WinSnoori and WinPitch are merely signal editors which can be used in conjunction with speech synthesisers. Here are some examples found in the literature:

- Germain and Martin (2000).
- Bonneau et al (2000) use of WinSnoori,

However, Germain-Rutherford (2001) describe Win Pitch LTL, a version of WinPitch which has been modified for language learning applications. Indeed, WinPitch LTL has specific/special functions for use in language learning applications:

- '*ralenti'* speech rate manipulation;
- '*traçage'* pitch tacking, stylisation of intonation patterns;
- '*texte'* permits annotation by teacher – e.g. can associate letters/phonemes to parts of the output;
- '*boites de commentaires'* can annotate and give feedback like written homework;
- *'cut/copy/paste'* e.g. get rid of aspiration when made where it shouldn't be and compare segments;
- '*synthesis'* re-define intonation etc and re-synthesise so that the student can hear their own voice;
- '*leçons'* the teacher can author their own lessons.

WinPitch is also used by (Fagayal and Golato, 2000).

Due to the fact that in most cases the developers are re-using speech synthesisers that were developed for use as reading machines (or for other dedicated purposes) and because most of the commercial synthesisers used are based on concatenation it is not possible to manipulate many of the parameters of the speech. (In general, concatenative systems only allow the manipulation of the following parameters within a small range speech rate, volume (amplitude) and pitch).

Keller and Zellner-Keller (2000) demonstrate how limited current commercial systems are by calculating rough estimates of the number of possible voices and styles. The number of voices and styles available is seriously restricted; in general only one voice and one style (typically reading) are available.
This is a severe limitation and it is unlikely to improve until more efficient methods (Automatic methods cannot be relied on yet) of creating new voices and styles are developed, for at the moment, in order to produce a new voice or style you need to collect a whole new speech database. In addition, emotional speech is still not yet truly possible.

Consequences of the re-use of speech synthesis systems: they are not necessarily adapted for language learning applications. Such systems should be evaluated prior to their integration in such a applications and, eventually and if possible (as for WinPitch LTL), adapted.

How will we know if the output is suitable for use in CALL applications?/What criteria should we use in our evaluation? According to Keller and Zellner-Keller (2000) "When the language competence of the system begins to outstrip that of some of the better second language users, such systems become useful new adjunct tools."

In fact, comes out of Bradlow et al. 1996 evaluation of TTS, an evaluation based on intelligibility, the following description:

> "such a talker would be a female who produce sentences with a relatively wide range in fundamental frequency, employs a relatively expanded vowel space that covers a broad range in F1 [this is language dependent], precisely articulates her point vowels, and has a high precision of inter-segmental timing." (p. 270)

So far the evaluation of TTS has not/rarely been done in the context of language learning applications. Most only indicate that they have conducted informal evaluations. Early examples of conducted evaluations are found in Sherwood (1981) and Stratil (1987). More recent evaluations are described in Santiago-Oriola et al. (1998) and currently in progress

(Handley and Hamel 2002). The latter highlight the importance not only of the accuracy and the reliability of the output but also its intelligibility as an essential requirement for such systems in such environments.

7 Conclusion

This article has presented a brief overview of the use of speech synthesis in CALL. It has many (potential) advantages/benefits to language learning. In particular it can be used to help familiarise learners with phoneme-graphme and graphme-phoneme relationships, to support pronunciation and to provide the learner with a dialogue partner. TTS gains when enhanced by visual output, articulatory one in particular, by the use of 3D talking heads more precisely. However, as most speech synthesisers seem to be borrowed and because language learning has some critical requirements, speech aynthesis for language learning and teaching purposes should be tested/evaluated.

References

Aist, G. (1999). Speech Recognition in Computer-Assisted Language Learning. Pennington?, Swets and Zeitlinger**:** 165-181.

Bonneau, A., Y. Laprie, et al. (2000). Towards Phonetic Tools for Speech Training. InSTIL.

Bradlow, A. R., G. M. Torretta, et al. (1996). "Intelligibility of Normal Speech I: Global and Fine-Grained Acoustic-Phonetic Talker Characteristics." Speech Communication **20**(2-3): 255-272.

Dutoit, T. (1997). An Introduction to Text-To-Speech Synthesis. Dordrecht, Kluwer Academic Publishers.

Egan, B. K. and S. A. LaRocca (2000). Speech Recognition in Language Learning: A Must. InSTIL.

Ehsani, F. and E. Knodt (1998). "Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of New CALL Paradigm." Language Learning and Technology **2**(1): 45-60.

Fagyal, Z. and P. Golato (2000). Using real time speech analysis software to teach the sounds of French. InSTIL.

Fatah, Elturan, et al. (1998). A Proposal for Teaching Kurdish Children their Native Language Skills.

Germain, A. and P. Martin (2000). "Présentation D'un Logiciel de Visualisation pour l'Apprentissage de l'Oral en Langue Seconde." ALSIC **3**(1): 61-76.

Germain Rutherford, A. (2001). How Speech Technology can Enhance Feedback in Teaching Oral Skills in a L2. **2002**.

Godwin-Jones, B. (2000). "Speech Technologies for Language Learning." Language Learning and Technology **3**(2): 6-9.

Hamel, M.-J. (1998). "Les Outils de TALN dans SAFRAN." ReCALL journal **10**(1).

Hamel, M.-J., B. Nkwenti-Azeh, et al. (1996). The Conceptual Dictionary in CALL. EUROCALL '95.

IRISA (2000). Projet CORDIAL: Communication multimodale personne-machine à composantes orales: méthodes et modèles. Lannion, France, IRISA (Institut de Recherche en Informatique et Systèmes Aléatoires.

Keller, E. and B. Zellner-Keller (2000). Speech Synthesis in Language Learning: Challenges and Opportunities. InSTIL.

Knoerr, H. (2000). "Pratique Intonative et Utilisation d'un Logiciel de Visualisation dans un cours de Prononciation en Français Langue Seconde: une Etude Descriptive." revue canadienne de Linguistique Appliqué **3**(1-2): 123-140.

Komissarchick, E. and J. Komissarchick (2000). Suprasegmental Pronunciation Training using Advance Knowledge-Based Speech Analysis. InSTIL.

Last, R. W. (1989). Artificial Intelligence Techniques in Language Learning. Chichester, Ellis Horwood.

Mercier, G., M. Guyomard, et al. (2000). Courseware for Breton Spelling, Pronunciation and Intonation Learning. InSTIL.

Moisa, T. and D. Ontanu (1999). Learning Romanian Using Speech Synthesis. Advanced Research in Computers and Communications Education, IOS Press.

Mumford, E. (1998). The FREDA project. CCL. Manchester, UMIST.

Myers, M. J. (2000). Voice Recognition Software used to Learn Pronunciation. InSTIL.

Nagano, K. and K. Ozawa (1990). English Speech training using Voice Conversion. International Conference on Spoken Language Processing (ICSLP), Kobe.

Osborne, M. (2000). Incorporating Text-to-Speech and Speech Recognition in Language Learning Applications using Microsoft Agent Technology in ToolBook II Instructor. InSTIL.

Riekki, V.-P. (1999). Paula Web Survey.
http://paula.oulu.fi/Publications/Survey/WebSurvey.pdf

Protopapas, A. and B. Calhoun (2000). Adaptive Phonetic training for Second-Language Learners. InSTIL.

Raskind, M. H. and E. Higgins (1995). "Effects of Speech Synthesis on the Proofreading Efficiency of Postsecondary Students with Learning Disabilities." Learning disability quarterly **18**(2): 141-158.

Santiago-Oriola, C. (1999). Vocal Synthesis in a Computerized Dictation Exercise. International Congress of Phonetic Sciences (ICPhS'99), San Fransisco.

Santiago-Oriola, C. and G. Pérenou (1999). From Grapheme to Phoneme: Diagnosis in a Dictation. ESUROSPEECH '99, Budapest.

Sherwood, B. (1981). "Speech Synthesis Applied to Language Teaching." Studies in Language Learning **3**: 175-181.

Siroux, Gourmelon, et al. (1998). "KGB Project: Tools and Resources for Breton Language Learning."

Skrelin, P. and N. Volskaya (1998). Application of New technologies in the Development of Education Programs. Language Teaching and Language Technology. S. Jager, J. Nerbonne and A. van Essen. Lisse, The Netherlands, Swets and Zeitlinger**:** 21-24.

Stratil, M., D. Burkhardt, et al. (1987). "Computer-Aided Language-Learning with Speech Synthesis: User Reactions." Programmed Learning and Educational Technology **24**(4): 309-316.

Stratil, M., G. Weston, et al. (1987). "Exploration of Foreign Language Speech Synthesis." Literary and Linguistic Computing **2**(2): 116-119.

Yoram, M. and K. Hirose (1996). Language Training System Utilizing Speech Modification. ICSLP, Philadelphia, USA.

Esling, J. H. (1992). Speech Technology in Applied Linguistics Instruction. Computers in Applied Linguistics. An International Perspective. M. Pennington and V. Stevens. Clevedon, England, Multilingual Matters Ltd**:** 233-272.

Gaudinat, A. and E. Wehrli (1997). "Analyse Syntaxique et Synthèse de la Parole: le Projet FIPSvox." TAL **38**(1): 121-134.

Hamel, M.-J. and A. Vandeventer (2000). Adapter un Analyseur Syntaxique et l'Intégrer dans un Système d'ELAO: le cas de FIPSgram dans SAFRAN. Enseignement-Apprentissage de la Langue Seconde dans des Environnements Multimédias. M. Laurier and L. Duquette. Montreal, Editions Logiques**:** 117-136.

LeGoff, B., T. Guiard-Marigny, et al. (1994). Real-time Analysis-Synthesis and Intelligibility of Talking Faces. Second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, New York, USA.

Micro-Video-Corporation (1990-96). VideoVoice Speech Training System ^{TM}, Micro-Video-Corporation.

Price, P. (1998). "How can Speech Technology Replicate and Complement Good Language Teachers to Help People Learn Language?" STILL: 81-90.

The-Speech-Institute (1995). Speech Assist Demonstration Disk, The Speech Institute @ Speech Institute Ltd.

Vandeventer, A. and M.-J. Hamel (2000). "Reusing a Syntactic Generator for CALL Purposes." ReCALL journal **12**(1): 94-104.

Bondarko, L. V., P. Skrelin, et al. (1996). RUSVOX-the Concatenation Speech Synthesis System for Russian. SPECOM'96, Copenhagen.

d'Alessandro, C. (2001). 33 ans de Synthèse de la Parole à partir du Texte: Une Promenade Sonore (1968-2001). Synthèse de la parole à partir du texte. C. d'Alessandro and E. Tzoukermann. Paris, ATALA/Hermès Science Publications**:** 297-321.

Hiller, S., E. Rooney, et al. (1994). "An Automated System for Computer-Aided Pronunciation Learning." Computer-Assisted Language Learning **7**(1): 51-63.

Auralog, Tell Me More – English http://www.auralog.com