

# Speech technologies and foreign language teaching and learning

Norbert Pachler

Institute of Education, University of London

“Has a panacea for FL teaching and learning at last been found?”

## INTRODUCTION

Information and Communication Technologies (ICT) are widely believed to have a useful contribution to make to foreign language (FL) teaching and learning. They can be used to enrich learners' FL learning experience and allow them to practise language skills independently of the teacher. Developments in multimedia technology enable the combination of authentic spoken and written texts as well as culturally rich visual stimuli supporting the constructivist learning paradigm, which is based on the premise that learning is a process of discovery, information processing and language generation. ICT give learners access to a range of resources in the target language and enable them to find out about how people of target language communities speak and live. Proponents of a social-interactionist view of foreign language learning stress the importance of the computer as tool for communication and (real-time) language use. (For an overview of ICT use in FL teaching and learning see e.g. Pachler and Field, 2001.)

This review article sets out to examine critically what one aspect of ICT, speech technologies, and in particular one popular program called *TeLL me More*, has to offer in relation to the potential of ICT in terms of practising learners' spoken language and listening skills.

## TELL ME MORE – THE PROMOTIONAL LITERATURE

*TeLL me more* is one of three different 'methods' on offer by Auralog<sup>1</sup>:

1. the conversation method: "Talk to me"
2. the key to success: "TeLL me More"; and
3. "TeLL me More Pro": the professional solution. (Auralog, no date)

The promotional literature sent out with the review copies of the CD-ROMs (Auralog, no date) makes considerable claims about the company's products

and is keen to point out that products two and three above have been approved for use by the French Ministry of Education for their educational value. Also, it stresses how the French Ministry has equipped all of its university departments for teacher training (IUFM) and local centres for educational resources (CRDP) with copies of product two. Furthermore, it is noted that all secondary schools in Andalucia, Spain, have been issued with the same package by their regional government, the Junta de Andalucia. In the commercial world, companies such as Mercedes-Benz, Air France and l'Union des Banques Suisses have apparently purchased a large number of licences for in-house training purposes. In the UK, the package also enjoys popularity in schools and universities<sup>2</sup>.

The language used by the promotional literature is quite striking: it is asserted that the integration of speech recognition technologies into software 'heralded the dawn of a new era' with students 'enter[ing] into a conversation with the computer, which analyses their pronunciation, evaluates it and even answers them!'; 'thanks to multimedia', the reader is told, 'perfecting a foreign language has never been so easy and interactive!' Apparently, the students are freed from the constraints of sensory perception and offered a scientific evaluation of pronunciation and the student is allowed to enter into 'a fluent, interactive dialogue with the computer'.

Has a panacea for FL teaching and learning at last been found?

## WHAT TECHNOLOGY CAN AND CANNOT DO

In order to appreciate the potential benefit and, importantly, understand its limitations and so resist the unwarranted claims of eager sales departments, it is useful to know what (speech) technology is and is not yet capable of.

Playing audio recordings via a computer (or devices such as CD/DVD/MP3 players) – as

opposed to a tape recorder – requires digital, rather than traditional analogue, recordings. Digital audio has a number of advantages over its analogue ‘predecessor’: first of all, recordings tend to be of much better quality and do not deteriorate over time and with use. Also, digital audio offers random access and variable playback speeds, both of which are very useful for FL teachers and learners. Often, audio recordings are not used in lessons despite the teacher’s best intentions, for example, because of the difficulty of finding the right spot on the tape in a hurry. Also, which FL teacher has not encountered protestations from their charges during listening activities that the near-native speech of a tape recording is too quick and hence de-motivating for them?

The recognition and understanding of human speech are highly complex endeavours and they require a vast amount of knowledge, *inter alia* about phonological, lexical, semantic, grammatical, and pragmatic conventions.

An important distinction to be made when talking about speech technologies is that between *speech recognition* and *speech synthesis*.

### Speech recognition

Speech recognition<sup>3</sup> basically refers to the computer’s ability to ‘understand’ human speech, for example, to record, analyse and re-play it faithfully, and also to act upon it, e.g. to convert spoken input into written output on a word processor or to activate certain functions in a program through voice activation etc. Currently a number of useful dictation products are available commercially allowing for so-called speech-to-text operations and even some handheld digital dictaphones have the option of speech-to-text output to computers in addition to simply saving recordings as sound files.

Humans and machines process speech in fundamentally different ways...Complex cognitive processes account for the human ability to associate acoustic signals with meanings and intentions. For a computer, on the other hand, speech is essentially a series of digital values. However, despite these differences, the core problem of speech recognition is the same for both humans and machines: namely, of finding the best match between a given speech sound and its corresponding word string. (Ehsani and Knodt, 1998: 47)

Needless to say, the route to the solution of the best match between speech sounds and their corresponding word strings is invariably different.

Voice input to programs such as the ones mentioned above do, however, tend to be conceived predominantly as productivity tools for use in office environments rather than as FL learning software for application in classrooms and formal learning contexts. Speech recognition/

speech-to-text programs are typically trained to recognise an individual user’s voice input rather than that of groups of people – which would, for example, allow their use for recording and processing (i.e. analysing by way of concordancing programs<sup>4</sup> etc.) pair- and groupwork – native speakers not struggling beginners, and they are not generally geared to processing of continuous speech but rather to be used in a controlled environment with a limited vocabulary.

According to Ehsani and Knodt (1998: 48) the so-called Hidden Markov Modelling (HMM)-based approach to speech recognition has proven an effective method for creating high-performance speaker-independent recognition engines that can cope with large vocabularies. The vast majority of commercial systems apparently deploy this technique, which consists of the following five basic components:

- a) an acoustic signal analyzer which computes a spectral representation of the incoming speech;
- b) a set of phone models (HMMs) trained on large amounts of actual speech data;
- c) a lexicon for converting sub-word phone sequences into words;
- d) a statistical language model or grammar network that defines the recognition task in terms of legitimate word combinations at the sentence level;
- e) a decoder, which is a search algorithm for computing the best match between a spoken utterance and its corresponding word string.

A crucial consideration for automated speech recognition use in FL learning revolves around the human-machine interface design as well as the recognition of the domain specificity of speech recognition performance, i.e. the ability of machines to do only what they are programmed to do. The effectiveness of an application depends crucially on the size and nature of the vocabulary it has been trained for. This is particularly crucial in relation to speaker-independent recognition of continuous dictation (as opposed to recognising spontaneous conversational speech). Therefore, performance can be enhanced by strictly limiting the vocabulary size and complexity of the performance domain, i.e. the average branching factor, as this requires less processing time and memory, as well as by maximising the acoustic quality of the input and training the system to the speaking style of the user. (See Ehsani and Knodt, 1998: 49-50.)

FL learners pose a particular problem in this respect as their enunciations and renditions will invariably be approximate and differ over time as their pronunciation improves. Furthermore, the pronunciation of speakers with different mother tongues will be characterised by different phonetic

---

**“A crucial consideration for automated speech recognition use in FL learning revolves around the human-machine interface design”**

---

qualities and differences in tonal quality, stress and intonation, duration and tempo etc.

For FL learners the quality of feedback both at segmental level (e.g. response latency, segment duration, inter-word pauses in phrases, spectral likelihood and fundamental frequency) as well as at supra-segmental level (e.g. intonation and stress, loudness, duration and tempo) is very important. Some programs offer graphical displays such as speakers' faces, the vocal tract, spectrum information or speech waveforms. Such displays can be seen to have the potential to provide formative feedback but they need to be accompanied by guidance and instructions on how to interpret the displays.

The use made of speech recognition in FL learning software is currently mainly in the context of word/phrase discrimination, e.g. for learners to be asked to read out words and compare them to pre-recorded native speaker examples in order to appreciate differences in sounds and (nuances in) pronunciation, and word order/syntax transformations, e.g. by incorporating speech recognition in question formation and answering, transformational grammar exercises, and responses to audio/video input. (See also Godwin-Jones, 2000: 7.)

By and large these programs use a closed response (as opposed to an open response) design type where learners must choose one response from a limited number of possible responses and have a very limited choice as to the nature of their response. Recognition accuracy depends crucially on task definition (i.e. closed versus open), vocabulary size (small versus large) and the degree of non-native disfluency. This way so-called query implementation remains simple and the vocabulary needed small, and such systems can, therefore, be quite robust, yielding accuracy rates towards 100%. In an open response design, systems also proceed as if learner responses were selected from a multiple choice list. As a minimum, therefore, all possible correct responses must be pre-programmed. If the system is also to provide formative feedback, potential mistakes need to be modelled and predetermined. Yet more advanced systems are user-adaptive, i.e. they construct an evolving model of the user's knowledge by keeping track of error rates and presenting subsequent material accordingly. In simulated real-life conversations each encounter has to be modelled as well. (See Ehsani and Knodt, 1998: 54-5.)

For FL learning contexts it is also important for technology to be able to understand a user's spoken input and evaluate it for appropriateness, not just correctness, and to diagnose users' problems with usage and not just pronunciation and syntax. Alas, at present speech recognition in FL software is still insufficiently advanced to do so reliably.

### Speech synthesis

Speech synthesis is the process of a computer

generating text-to-speech, i.e. to electronically produce speech, or even to interact by way of speech with a human interlocutor, i.e. to talk back. One useful application of speech synthesis in FL learning is the use of programs such as *Read Please*<sup>5</sup>. This free program allows the user to have any English text read out to her by the computer. With this and similar programs it is possible to choose between different types of 'speakers' and to adjust the speed of rendition. There are also web-based applications such as the one offered by Bell Labs, an online multilingual text-to-speech speech synthesis tool<sup>6</sup> or Webspeech<sup>7</sup> which can be integrated into a web-browser.

### Reading aloud tutors

Another potential application of speech technologies are programs that teach learners how to read FL texts by 'listening' to learners' renditions and by providing feedback by way of intervention, correcting mistakes or providing help. Whilst designing a system where there is only one correct spoken response to any given written prompt appears to be relatively straightforward, according to Ehsani and Knodt (1998: 53) a considerable technical challenge lies in recognising and responding adequately to the disfluencies of inexperienced readers such as hesitations, mispronunciations, false starts and self-corrections.

## LEARNING THEORIES, CURRENT DEVELOPMENTS IN FL METHODOLOGY AND THEIR IMPLICATIONS FOR SOFTWARE EVALUATION

### Some theoretical considerations

Despite the prevalence of communicative competence in FL teaching and learning, traditional, behaviourist modes still tend to be very common in the world of CALL. They are characterised by breaking learning down into small manageable steps as well as drill-practice following the 'presentation-practice-production' model rather than, for example, one of 'observe-hypothesise-experiment' (see Lewis, 1993).

A constructivist approach to FL learning, on the other hand, posits that

- learning must be regarded as an active and collaborative process of knowledge construction;
- learning is to be seen as an autonomous process, to be regulated by the learners' expectations, goals, existing schemata and intentions;
- learning is a process of experimentation based on previous knowledge and experience;
- learning is a process of socially negotiated construction of meaning;
- learning is a process which must be

**“behaviourist modes still tend to be very common in the world of CALL.”**