An Introduction to Speech Technology in Language Learning

Table of Contents

One of the most technically advanced and most exciting area of development in CALL today is the use of speech technology. It is also particularly important as speech interaction is an essential ingredient in the prevalent communicative approach to ab initio language learning. Simply doing written exercises on the computer screen will not do, learners must listen to authentic speech and reproduce it.

Speech Technology in Language Learning henceforth referred to by the acronym STiLL includes four main thematic components which will be studied consecutively in this module, speech recognition, speech synthesis, audio-visual synthesis including talking heads, and visualisation.

1. Speech recognition: the following section represents up to **5 hours of work** including the practice assignment.

1.1 What is speech recognition?

Recognition is often referred to in the literature as Automatic Speech Recognition (or ASR). In lay man's terms, speech recognition can be defined as "the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words." (Zue, Cole and Ward, 1996) Once recognised, the words can be used as input to any number of different applications. The recognised words can be used to control computers or other machines, for data entry and for text processing.

1.1.1 How does the recognition process work? Speech to Text

Text processing is probably the best known application of speech recognition following the launch of dictation products such as IBM's ViaVoice and Dragon's Naturally Speaking. As people will be more familiar with these products than other speech recognition products, this is the best starting point for understanding how the technology works.

In this context, recognition is simply the process of converting an acoustic signal into a transcription of the speech sequence, a process analogous to doing a dictation. Consequently, this type of recognition system is often called a speech-to-text system.

In general, the process of dictation can be broken down into the following tasks:

1. Data capture
2. Pre-processing
3. Recognition or pattern matching
4. Hypothesis filtering
5. Transcription

The speech signal is first captured using a microphone or a telephone. This analogue signal is then converted into a digital signal. Audio files can be very large and contain a lot of information. However, we know that a lot of this information is redundant/unnecessary for the identification of speech sounds. Although a human ear can theoretically hear sounds within the range of 20 to 20,000 Hz, for speech purposes, it is most sensitive to sounds within the range of 1,000 to 6,000 Hz. Consequently, most recognition systems incorporate a pre-processor which either takes samples of the speech signal or only extracts the parameters that are believed to be necessary and sufficient to enable accurate speech recognition.

The speech signal is then divided into smaller segments, the length of which can vary from a phoneme (individual distinctive speech sound as in phonetics/linguistics) to a complete sentence depending on the system employed. Next, the recognition module attempts to match these segments to segments stored in its database. Several different methods are used:

- Pattern-matching
- Acoustic-phonetic recognition
- Stochastic (or statistical methods)
- Artificial neural networks (a completely different process from the statistical matching techniques mentioned above)

1.1.2 Pattern matching

This technique, which was developed in the 1950s, consists in the creation of a database of reference templates of the chosen minimal units (usually words). In order for the computer to recognise the speech input, it compares the pattern of the uttered words with all the stored templates and tries to find the best match. To stop ambient noise being incorrectly recognised as speech, most systems use distance metrics (thresholds of acceptance), these ensure that utterances which are not close enough to the stored templates are not transcribed. This might also be called a global or word based approach. It is also known as "template-based speech recognition".

1.1.3 Acoustic-phonetic recognition (also called phoneme based ASR)

The creation of an acoustic-phonetic recognition system begins with the construction of a database of reference models of the different phonemes of the language in question. The reference models are composed of the features that are necessary and sufficient to distinguish one phoneme from another. These features are obtained by experts through the analysis of the spectral patterns of the different phonemes of the language in question. Recognition is again a process of pattern matching. Although phonemes are identified, matching takes place at the word level and for each word, we obtain a set of phoneme hypotheses in the form of a trellis. These hypotheses are then compared with the words stored in the systems lexicon and the best match is chosen.

1.1.4 Stochastic (statistical) methods

Stochastic methods are essentially pattern-matching which incorporates statistics to "select the highest probability outcome from a sequence of samples." (Speech Recognition – The State Of The Art). The most frequently used statistical model in speech recognition is called the **Hidden Markov Model (HMM)**, which is "a numeric analysis which determines the probability of the next item in a string of items" (21st Century Eloquence Speech Recognition – Glossary of Terms).

The use of HMMs in speech recognition was proposed because there are restrictions on the combination and distribution of elements in a sentence. The realisation of a sound depends on its context; not all sounds have the same probability of appearing at the beginning of a sentence, the sounds which follow are conditioned (depend on) the sounds that preceded them. (Dubois et al, 1999). To give an example, we could cite consonant clusters (groups of consonants) in English, if you start with the sequence ST, it can be followed by very few consonants, so the consonant R would be highly predictable as in the word STRUCTURE.

1.1.5 Artificial Neural Networks (or ANNs)

ANNs are essentially computer models that attempt to emulate the way in which the brain processes language. Like the brain, they do not need to start out with a complete definition of the problem, because they learn through exposure to data. This a promising area but it is not used yet in commercial applications in ASR. Neural Networks are used in many other forms of advanced programming.

Today commercial systems are based on HMMs. Pattern-matching is sufficient for small vocabularies, however it cannot necessarily make the fine distinctions necessary to process larger vocabularies. ASR then becomes accurate to a ratio of 90% for the best systems, this is still very helpful. Acoustic-phonetic recognition is labour intensive; many hours of manual work must be done in order to acquire the necessary linguistic knowledge. On the other hand, systems based on HMMs can be trained automatically. They are fast, efficient and robust. The only disadvantage of HMMs is that they cannot precisely model the creative nature of language; they are only really suitable for applications in restricted domains (particularly where the field of answers is easy to anticipate). Consequently, today, much time and money in beyond state of the art research is being invested in ANNs because the networks can automatically learn new combinations.

None of these methods is 100% accurate, therefore many speech recognition systems incorporate an additional "linguistic" module which serves to filter and validate the hypotheses identified by the previous module. This supplementary linguistic information may be theoretically motivated and based on a traditional linguistic theory or it may be based on a statistical model. The most frequently used statistical models are n-grams. **The idea behind n-grams is that the appearance of a word depends on the "n minus 1" words that have preceded it**.

Finally, the transcription of the identified words is obtained from the reference lexicon.

Computers are not yet sophisticated enough to actually understand freely produced utterances, consequently most contemporary control-command and data entry systems are mainly based on pattern matching (matching sound sequences to words in the database vocabulary). Users are expected to use set phrases when communicating with the computer.

Before going on to discuss some of the products currently available on the market, it is necessary to present the factors which distinguish them. These factors, which should be taken into account when acquiring a speech recognition system, are:

- mode of speech input (discrete vs continuous speech)
- speaker dependency
- size of vocabulary
- intended domain of application

1.1.6 Mode of speech input: recognising **separate** (discrete) words against recognising the **continuous** flow of speech.

In general, a dichotomy exists between systems which can only process **discrete speech** and systems which can process **continuous speech**. When using a system that supports discrete speech or isolated speech, **the user must pause between each pair of words**, this pause must be distinct from the pause that is used to mark the end of a sentence. In fact mode of speech input is not a pure dichotomy. In between the two extremes systems can be found which recognise connected speech and others which can **spot individual words in continuous speech**. These are called **word spotting** systems. Systems which support connected speech do not require the user to pause between pairs of words, however users must not insert inter-word articulatory effects such as liaisons when using French systems. (Rodman, 1999). This is because otherwise word recognition is lost. This phenomenon is relevant to syllable based spoken languages such as French. Word stressed languages such as English offer much clearer distinctions between words.

1.1.7 Degree of dependency on the speaker (speaker dependent vs speaker independent)

Traditionally, a distinction is made between **speaker dependent** and **speaker independent** systems. Speaker dependent systems do not take into account inter-speaker variability. They generally only recognise the voice/speech of one individual user and are difficult to adapt for use by other speakers. Speaker independent systems model speaker variability and can therefore accept speech input from a number of different users.

Paradoxically, many commercial systems are initially speaker independent, but over time they dynamically adapt to the voice of the user and consequently become increasingly speaker dependent. (Rodman, 1999)

This is a crucial distinction which has led to the terminological difference between **Voice Recognition** which is speaker dependent and **Speech Recognition** which is speaker independent. Voice recognition implies training the machine to recognise its master's voice. Speech Recognition involves recognising many different speakers. It is easy to remember this difference by remembering that voice is meant to be an individual characteristic of a speaker whilst speech is a universal ability of the human species.

1.1.8 Vocabulary size

Intuitively, we would be inclined to assume that the larger the vocabulary, the more likely the system would be to commit errors due to the number of similarly pronounced words. The number of entries in the systems vocabulary is important, however the choice of lexical items is often more important. Even a small vocabulary such as one for the letters of the alphabet can cause recognition errors due to the number of phonemes and minimal pairs (Rodman, 1999). The choice of lexical items is also important with regard to storage requirements. In order to reduce storage requirements when treating languages like Spanish which have an extremely rich inflectional morphology the systems lexicon should at least contain the canonical forms (roots) of each set of derivations, unpredictable derivations (irregular) and a method of predicting the pronunciation of regular derivations.

Practice Assignment 1: test your knowledge of ASR

Preferably without referring back to the 4 pages above, attempt to answer the following questions:

1. What is the most popular method for achieving robust (reliable) speech recognition?
2. Give one well-known example of what recognition was initially used for?
3. Explain the difference between pattern matching and phoneme based recognition.
4. Out of the various forms of recognition, which is the odd one out (not yet ready for the market)?
5. How do statistics play a part in the recognition process?
6. Explain the difference between discrete and continuous speech recognition.
7. Explain the difference between Voice and Speech Recognition.
8. In what way does vocabulary size matter?
9. Can you guess what happens if your pronounce an item which is not in the speech database?
10. Does anticipitation improve recognition accuracy?

Please note: if you cannot answer more than 50% of these, you should read the section again making notes.

Answers:

1. Hidden Markov Models or HMMs
2. Speech to Text Dictation
3. Pattern matching is more global recognising word size entities whereas phoneme based recognition as the name indicates recognises individual distinctive speech sounds.
4. Artificial Neural Networks or ANNs are the new form of speech recognition technology.
5. In the recognition process, stochastic (statistical) techniques are used to match units or words to speech segments.
6. Discrete speech recognition relies on speakers enunciating words one by one with pauses, continuous speech recognition is more advanced as it recognises the normal flow of speech, although care has to be taken with liaison phonemena which blur word distinctions in languages like French.
7. Voice recognition is speaker dependent, the machine has to be trained to recognise "its master's voice", speech recognition is speaker independent and will recognise the speech of many different speakers.
8. The more words a speech recognition vocabulary has, the more language it is capable of recognising accurately, this is even more important for more open context recognition where the anticipated input/response is not known.
9. If an item does not belong to a speech database, the system has two ways of dealing with it which may be pre-programmed, either, it will reject the item and ask for it to be repeated or it will recognised the closest match.
10. Anticipation considerably improves recognition accuracy, indeed, in the current state of the art, it is fair to say that great accuracy is achieved by restricted the field of anticipated input or response.

1. 2 Why use speech recognition in language learning? The following section represents 6 hours of work including the two assignments provided.

The integration of speech recognition into the interface of CALL applications is particularly interesting/useful/important because it permits the transition from a passive to an active language learning environment. Before the introduction of speech recognition, CALL courseware offered the learner a variety of playback modes, yet the learners' interaction with the system was restricted to the use of the keyboard and the mouse. (Aist, 1999)

Until the introduction of speech technology, in particular speech recognition, to CALL activities, pronunciation training “was limited by the subject’s hearing abilities,” (Sioufi, 2000) that is to say by their perception. If a learner cannot perceive the difference between two sounds, even if they are aware that there is a difference between two sounds they will not be able to produce this difference. …

1.2.1 Integrating speech recognition into CALL activities – the challenge

A general purpose speech recognition system must overcome many challenges. The main problems they must overcome are:

- the huge quantities of data that must be processed
  For example, speech sounds can span a range of more than 20,000 frequencies. (Marcowitz, 1996)
- the difficulties of processing continuous speech
  Unlike the written word, spoken language forms a continuous flow; there are no explicit boundaries between the individual words analogous to the spaces between written words. This problem is accentuated by the presence of inter-word articulatory effects such as liaison etc.
- articulatory effects
  Without knowing the context of an utterance, articulatory effects such as assimilation (where one adjacent sound is affected by another) can make it extremely difficult for a speech recognition system to determine what has been said; articulatory effects can sometimes deform utterances to such an extent that they hardly ressemble the initial utterance.
  For example:
  Geechet? = Did you eat yet? (Rodman, 1999)
- inter and intra speaker variation
  It is not surprising that different speakers pronounce words differently and speak at different rates and so on and so forth. Physical factors such as the size of the speakers head have a direct effect on the size of the resonators and therefore on the frequency of the sounds they produce. Other factors which cause inter-speaker variation are the age, gender, dialect and ideolect of the speaker. What may surprise you though is that even an individual speaker never pronounces the same word in the same way; if they are enthusiastic, they speak faster; if they are angry, they speak louder.
- noise interference
  Non-speech sounds can hinder speech recognition if they are misrecognised as words. The main sources of noise when using speech recognition systems are:
  - ambient noise
  - the speaker
    - for example the speaker may sniff, cough, or yawn.
  - the method of data capture
    - for example the microphone

It is important to note that the word Noise is used technically in communication theory to refer to non intelligible utterances.

1.2.2 State-of-the-art of speech recognition in CALL

Speech recognition has been widely exploited in language learning. Experiments have been conducted on the use of dictation systems and control and command systems in CALL. In addition, several speech recognition systems have been specifically designed for the purpose of language teaching. The vast majority of this research has been focused on the development of courseware for conversation simulation and pronunciation training. In addition speech recognition has also been used for training in reading aloud and in simple dialogue systems.

Let us begin by considering how general purpose dictation software can be used in language teaching. Myers (2000) suggests that continuous dictation systems such as Via Voice Gold by IBM or Dragon Naturally Speaking by Dragon Systems Inc. can be used by language learners to write their texts in free production exercises. She highlights the following advantages of using dictation software for free-production exercises:

- Using dictation software makes learners more aware of sound patterns structured in phonetic segments. "This is because they must articulate clearly if the machine is to pick up words to the utmost level of accuracy." (Myers, 2000)
- The software allows learners to be more creative because they do not have to concentrate on the form (especially the spelling of what they are saying). In addition, because they do not have to concentrate on the form, learners are more likely to speak spontaneously rather than using set phrases and pre-fabricated syntax.

There are however some disadvantages to using dictation software that is intended for use by native speakers in language learning. Dictation systems and pronunciation training systems have entirely different objectives. "A conventional speech recogniser is designed to generate the most charitable reading of a speaker's utterance." (Ehsani and Knodt, 1998) "The focus is on recognition rather than rejection or measurement" (Egan and LaRocca, 2000). A conventional dictation system aims to identify what the user intended to say even when this deviates quite radically from standard pronunciation. With a pronunciation tutor on the other hand, the emphasis is on identifying exactly what has been said, not what the learner intended to say. A pronunciation tutor must be trained to recognise deviations form standard native pronunciation and must be able to give corrective feedback. In addition, "to be most effective for language learning, continuous speech recognition must be able to handle spontaneous speaking styles." (Ehsani and Knodt, 1998). Currently they can only really process "careful or planned speech." (Ehsani and Knodt, 1998) This, however, is not entirely a bad thing, it could be said that dictation systems provide a compromise between tasks with a communicative focus and tasks whose focus is on mastering linguistic form; to guarantee recognition learners must use standard speech therefore they must concentrate on the linguistic form of what they are saying, however some cognitive load is freed as they no longer have to concentrate on how to spell what they are saying, consequently they can focus more of their attention on the communicative task and be more creative.

Aist (1999) suggests that "it is unclear whether the accuracy of continuous speech recognition systems is high enough yet to support recognition except in cases where there are strong expectations of what the student will say."

Another advantage – consolidation and reinforcement of the association between written and spoken forms.

Assignment 2: Brainstorm

Before studying the various existing uses of speech recognition in CALL, write as many examples of what you could imagine using speech recognition for with your learners. This brainstorm should result in a number of bullet points such as:
- conversation simulation
- vowel training

Possible answers (brainstorm points):

- learning individual speech sounds
- work on minimal pairs (but hard for the recogniser)
- work on sounds which do not exist in one's own (source) language
- sentence production
- fluency training
- conversation simulation

1.2.3 Voice Recognition, Dictation and Interface Control

If you are thinking about experimenting with dictation systems in your language classroom, **dyslexic.com** offers some useful advice. In addition to the features which distinguish the different dictation systems discussed above, namely mode of speech input, speaker dependency and vocabulary size, when purchasing a speech recognition system you should also find out whether the system provides **speech feedback**, in other words whether it is coupled with a **text-to-speech system** (synthetic voice that reads text), and how much training the system requires. Speech feedback, is extremely useful for proof reading documents. "Proof reading, especially from a computer screen, is a difficult skill. We all tend to read what we want to read, rather than what is actually written there." With regard to the mode of speech input, people tend to assume that systems which support continuous speech input are naturally better, however some people "consider that discrete speech is better, particularly for young children, as the cognitive load in handling dictation is easier one word at a time. … **DragonDictate** is the only discrete speech program still fully available, but Dragon have stopped work on it. They argue that you can dictate satisfactorily in separate words with the continuous speech products if you want to." (Dictation [Voice Recognition or VR] systems compared).

A presentation of the main features of these systems and others recommended by dyslexic.com, including those produced by IBM, L&H and Philips, is available at the following address: http://www.dyslexic.com/dictcomp.htm It is also possible to purchase the software from this site.

To conclude here is a brief summary of the major differences between the software produced by these companies. IBM Via Voice is the only speech recognition system currently available for the Apple Macintosh. Naturally Speaking requires 5 minutes of training before use, Via Voice Millennium on the other hand requires 16 minutes. Via Voice Millennium is particularly difficult to train for learners as it requires you to read

quite long paragraphs. The training phase for L&H Voice Express is even more difficult. As far as accuracy is concerned, opinion is divided. Some believe that Dragon Naturally Speaking makes the least errors, others believe that IBM Via Voice gives the best accuracy. Most of the products produced by IBM, Dragon, and Philips provide speech feedback. L&H Voice Express unfortunately does not. L&H Voice Express can be integrated with Word. IBM Via Voice Pro has the advantage that it can be used to voice enable not only Word, but also Excel and Outlook.

Please note that, at the time of going to press, the major Language & Speech company, Lernout & Hauspie (L&H), placed into receivership in 2001, has been acquired by US company Scansoft, it is not yet known what impact this will have on the availability of the L&H range of products.

Due to the huge challenges involved in speech recognition, and in particular the recognition of speech produced by learners, command and control systems were the first examples of speech recognition software to be integrated into CALL applications. Command and control systems were used to provide the learner with an alternative to using the keyboard or the mouse to navigate between the different pages of the courseware, for instance, with the command, "Open Word" or "Open Excel"... TraciTalk The Mystery (available from Clarity Language Consultants Ltd. http://www.clarity.com.hk/ or Courseware Publishing International Inc., CA) is an example of CALL courseware that uses speech recognition for this purpose. “Critics of speech integration in language learning use this application to declare that automatic speech recognition is not ready for CALL” (Bonaventura, Howarth and Menzel, 2000). To a certain extent this might be true, but the usefulness of such applications must not be overlooked. “Speech is the most human interface with machines.” (Bonaventura, Howarth and Menzel, 2000). Such simple uses of automatic speech recognition, give learners a chance to build confidence in communicating in the target language.

1.2.4 Pronunciation training

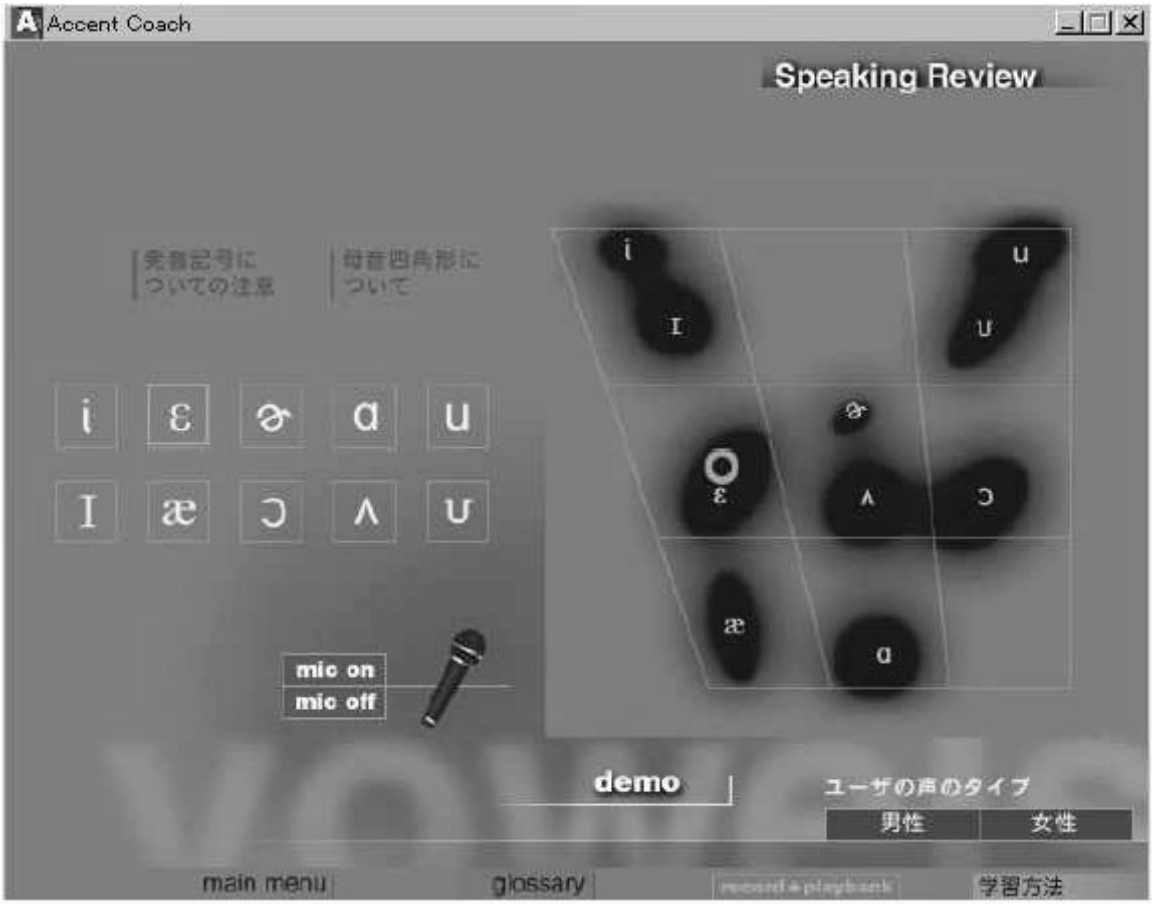
Several different methods of pronunciation training have been experimented with over the years. These include:

- visual feedback
- pattern matching
- error modelling

1.2.4.i Visual feedback (this is studied in more depth at the end of the module)

Several types of visual feedback have been used in CALL applications. The first and most obvious technique used was to present the learner with a graph of the waveform of the utterance they had produced alongside one of the same utterance produced by a native speaker.
Other more complex methods involve extracting certain features such as **fundamental frequency (essential tone information)** from the acoustic signal and presenting these

features to the learner in the form of simpler graphs. For example, this technique has been used to give feedback on **intonation**. The intonation of a phrase is presented to the learner using **pitch tunnels**. A pitch tunnel, is a graph representing the pitch values that the prosodic curve of their utterance should fall between. To provide feedback, the pitch contour of the utterance that the learner has produced is superimposed onto the pitch tunnel to show the learner where they have gone right or wrong. (Hiller et al, 1994). Visual feedback has also been used in teaching vowel production. Hiller et al. (1994) describe the use of **vowel targets**. Vowel targets are templates, derived from a corpus of native speakers utterances, which represent the parameters between which the vowel produced by the learner should fall. The learner receives visual feedback in the form of a graph indicating where the learners vowel fell in relation to the vowel targets. Some very attractive modes of visualisation were produced in this regard, for instance, darts or sprites were shown to hit a vowel target in real or near real time, this very entertaining method can be used to show how accurate or inaccurate learner vowel production can be. The additional advantage is that neighbouring sounds can be placed on the same target to illustrate possible and perhaps typical foreign learner errors. The figure below shows **Accent Coach** produced by Syracuse, for more details see the full review of the product shown on the CALICO reviews web site at
http://astro.ocis.temple.edu/~jburston/CALICO/review/accent00.htm



The above figure shows the **Accent Coach Interactive Vowel Chart** used here by Japanese learners of US English. The **green circle** shows that the subject has pronounced the mid front vowel found in the word "**set**".

Such visual feedback is especially useful when teaching learners with hearing difficulties. However, to be effective systems based on the use of spectral patterns for feedback must provide the learner (and the teacher) with detailed explanations of how to interpret them. (Ehsani and Knodt, 1998)

1.2.4.ii Pattern matching

Pattern matching has been used to compare learners utterances with those of native speakers. In general, the user receives feedback in the form of a score indicating how close their pronunciation was to that of a native speaker.
(Such systems are not too difficult to create. It is relatively easy to adapt a conventional speech recogniser for this type of use in pronunciation training because most speech recognition systems incorporate distance metrics to ensure that noise is not misrecognised as words and so on and so forth [Egan and La Rocca, 2000]). French company **Auralog**, leading player in speech recognition driven courseware use a recogniser capable of scoring on a scale of 1 to 7 from very poor non native (1) to excellent native (7).

The disadvantage of pronunciation training systems that are based purely on scoring is that they do not provide adequate feedback to the learner. Learners often know that they are making mistakes, however they may not be able to exactly pinpoint the source of the mistake or know how to remedy it. ("Non-diagnostic feedback … is frustrating for students who don't know where they are making their mistakes." (Aist, 1999)) For this reason, systems which score units smaller than the sentence, for example words or even individual phonemes, are more useful than those which score whole sentences, because they enable the learner to locate their errors. However, these systems are still not ideal as they do not indicate to the learner how the problem can be remedied.

1.2.4.iii Error modelling (expert systems)

These are examples of expert systems based on knowledge of common errors (mispronunciations) made by language learners. These systems make the most of the current state of the art in speech recognition by simplifying the recognition task. The main component of these systems is a set of rules which model the errors that learners are expected to make when learning the target language. "For example, given the English word 'thin', common errors by native speakers of Italian may be modelled as 'tin', 'sin', and 'fin'." (Aist, 1999) "By modelling phonetic substitutions, pronunciation courseware can try to determine what the student said – not just how closely the utterance matched a template or acoustic score." (Aist, 1999) In other words, it can not only determine when a sentence has been correctly pronounced, but also when an error has been committed, it can locate it and then by identifying the rule that was applied to produce the error, it can give feedback to the learner on how to improve their pronunciation.

Although very simple, such systems have proved very useful in pronunciation training. However, they do have a few disadvantages:

- Although they are very good at treating common errors, they are not very good at detecting rare and more idiosyncratic pronunciations. (Ehsani and Knodt, 1998)
- More often than not these systems are based on language pairs and are difficult to re-use/adapt to other language pairs.
- They also assume that the phonetic system of the target language can be accurately mapped to the learner's native language. Mapping is generally only possible if the languages involved are from the same family (are related).
- They are not very good at measuring improvement; in general they can only tell the learner if what they said was right or wrong. Once the source of the error has been identified, a score based method may be better at detecting improvement. (Aist, 1999)

1.2.4.iv Training in reading aloud

Speech recognition has been used for reading training in both first and second language learning. Automatic speech recognition systems are particularly useful in the combat against illiteracy. "Illiteracy has stigmatizing status. The attraction of having such a tool on a computer is that the computer is precisely not a human, and is thus perceived as non-judgemental by illiterates." (Keller and Zellner-Keller, June 2000)

Examples of research projects include those conducted by the Center for Teaching and Learning (CTL) and project LISTEN at Carnegie Mellon University (CMU). Commercial software includes the Talking Books series from Sherston Software, UK, Let's Go Read from Edmark, Watch Me Read from IBM and Reader Rabbit from The Learning Company (TLC). (Aist, 1999)

Although the task of speech recognition is simplified because "the system 'knows' in advance what the student will be trying to say" (Ehsani and Knodt, 1998), it is still complex because the system must be able to deal with disfluencies including hesitations, mispronunciations, false starts and self-corrections (Ehsani and Knodt, 1998) and it may initially be uncertain where the learner is in the text. Consequently, early examples of reading training systems such as Let's Go Read were limited to phoneme drills and single-word exercises. Today's systems on the other hand can listen to children/learners read entire sentences and stories aloud.

When selecting a reading trainer teachers should look for a system which allows you to adjust the threshold of acceptance. In the beginning, learners will become frustrated if the system picks up on every little error including false-starts, self-corrections and hesitations. It may also be useful to look for a system which incorporates algorithms for measuring reading fluency and general performance.

1.2.4.v Teaching linguistic structures and limited conversation

Developments in speech recognition technology mean that speech recognition can now be used to "offer practice in a variety of higher-level linguistic skills ranging from highly constrained grammar and vocabulary drills to limited conversational skills and simulated real-life situations/conversations.

A general distinction is made between closed- and open-response systems.
Closed-response systems are essentially multiple-choice systems; learners must choose their response from a set of pre-defined responses.
With open-response systems the learner must "generate the appropriate response without any cues from the system." (Ehsani and Knodt, 1998)

Closed-response systems are naturally the easiest to design; they avoid many of the challenges posed to speech recognition because the system 'knows' what the learner is going to say in advance.

Closed-response systems are not restricted to simple multiple choice applications. If "the query-response mode is highly contextualised", it can be presented as part of a simulated conversation with a virtual interlocutor" (Ehsani and Knodt, 1998) Traci Talk The Mystery is an example of a CALL system that uses a closed-response system to provide simulated conversations. The system will be described in more detail below.

Open-response systems do not necessarily have to be much more complicated than closed-response systems. When used in very restricted contexts, for example for question-answer drills where the response is highly predictable or constrained, they may process students' responses in exactly the same way as closed-response systems, that is as if the response "were selected from a multiple-choice list." (Ehsani and Knodt, 1998) The Auto Interactive Tutor (TAIT) by Mitsubishi Research Laboratories is an example of a restricted context open-response system aimed at learners of Spanish (Ehsani and Knodt, 1998). The advantage of such systems over closed-response systems is that they provide a greater challenge to the learner.

On the other hand, open-response systems can be more complex. Examples include systems that attempt to simulate real-life conversation. The ultimate goal is to "build systems that can understand and judge continuous spoken language and maintain a conversation through several turns." (Ehsani and Knodt, 1998) These applications are rather more ambitious, however successful prototype systems have been developed. An example is Subarshii developed at Entropic. This system aims to give beginners of Japanese "the opportunity to solve simple problems through (virtual) spoken interactions with monolingual Japanese natives" in a constrained context (Ehsani and Knodt, 1998). The results of user trials have shown that simple dialogue systems may not be as difficult as we imagine to design. "Near perfect recognition accuracy may not be a necessary requirement for an effective speech dialog system" (Ehsani and Knodt, 1998).

1.2.5 Commercially available CALL products that integrate speech recognition

In 1991, **Auralog** were the first company to launch a CALL application based on speech recognition technology. We will therefore begin by presenting their product as they, without doubt, have the most experience in the domain. The Auralog web site also features a **History of Speech Recognition** which explains the process of using such technology in language learning.

Today, Auralog (http://www.auralog.com) produces 3 different types of CALL courseware based on speech recognition technology. These are:

- Te*L*L me More The Complete Solution
- Talk to Me The Conversation Method
- Te*L*L me More For Kids!

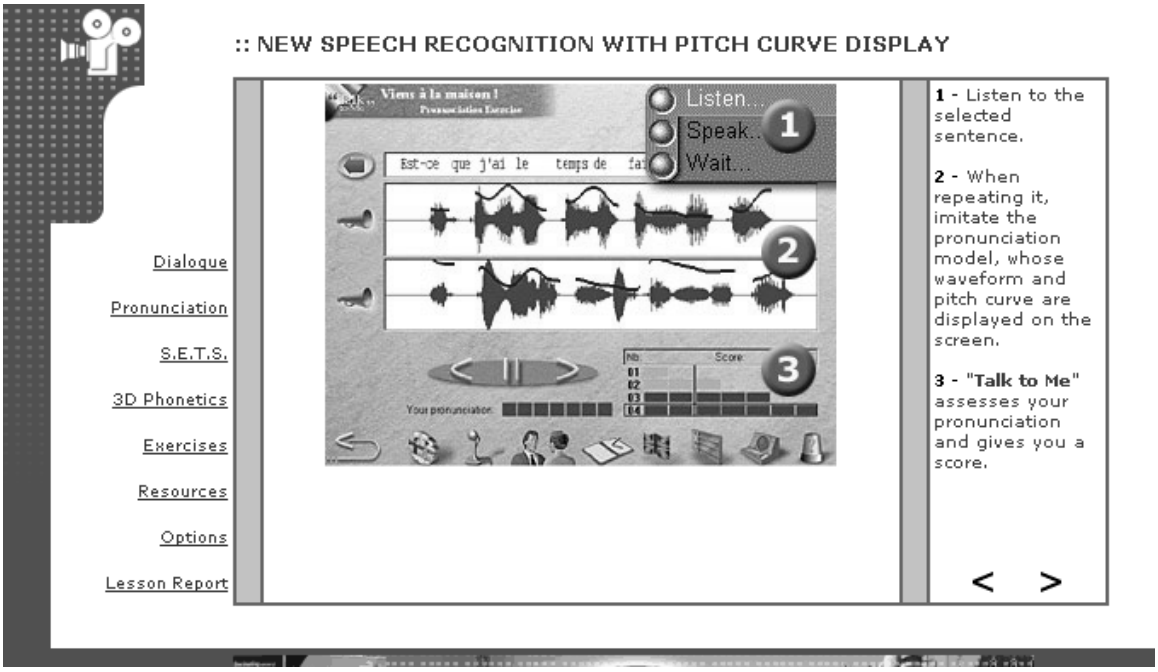
Prices start at $79. The software can be run on any machine with Windows installed. In general the software comes with its own headset so you do not have to worry about obtaining the right one.

Te*L*L me More The Complete Solution is available in 7 languages (Chinese, English, French, German, Italian, Spanish and Japanese) in four different levels of difficulty (beginners, intermediate, advanced and business which equate to GCSE level, A-level, degree and advanced degree level respectively). Talk to Me the Conversation Method is available in two levels of difficulty (beginners/intermediates and advanced). The major difference between these two packages is that Te*L*L me More The Complete solution focuses on non-communicative tasks such as grammar drills and vocabulary acquisition and pronunciation training based on imitation, whereas Talk to Me concentrates on communicative tasks through the use of dialogues (conversation simulation).

Interactive dialogues (screenshot obtained from http://www.multilingualbooks.com)

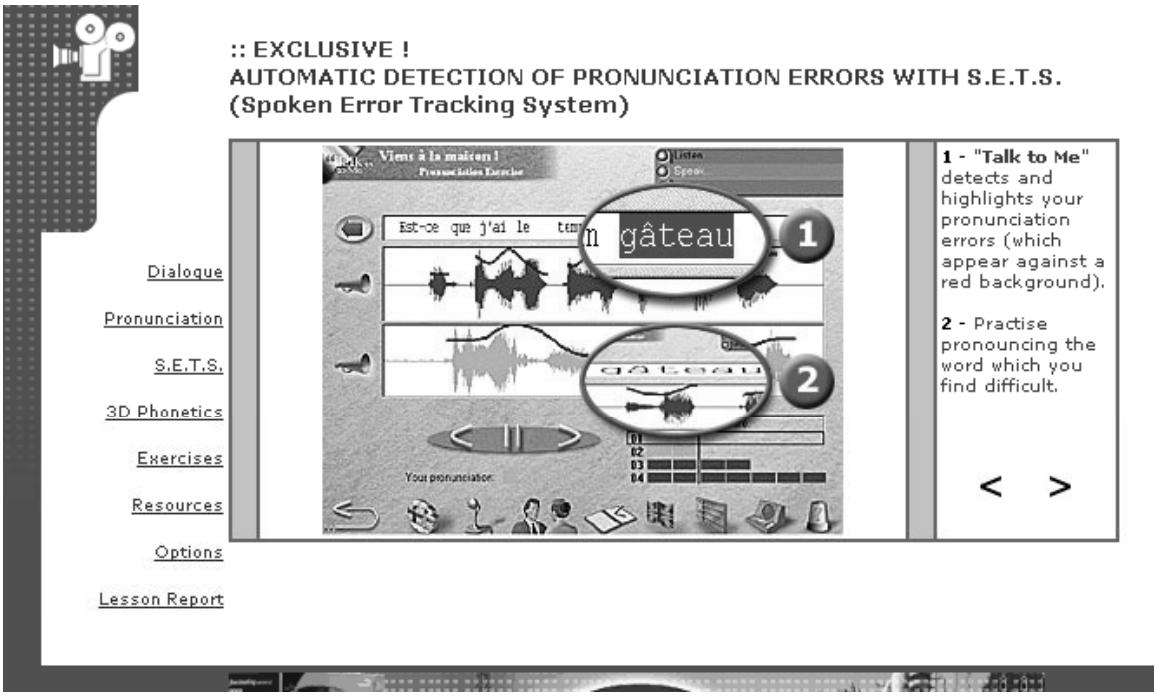
To give you an idea of what you are getting for your money, Te*L*L me More The Complete Solution provides you with 200 hours of language learning and over 1000 exercises.

Despite the difference in teaching methods they use essentially the same speech recognition technologies. Both score learners utterances and provide visual feedback. Two types of display mode are provided (waveform and pitch contour) which can be compared with a model produced by a native speaker.



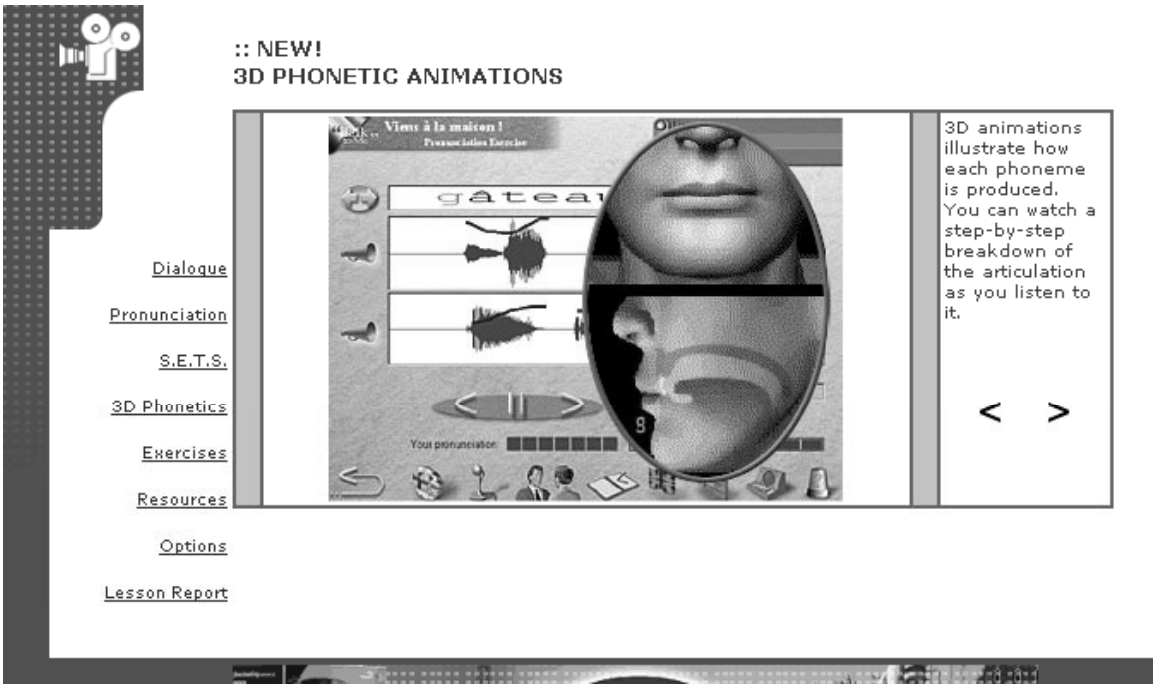
Waveform displays (Screenshot obtained from the Auralog web site)

In addition, the packages incorporate Auralog's **Spoken Error Tracking System (SETS)** which "pinpoints and highlights mispronounced words within individual sentences" (http://www.auralog.com/en/tellmemore.html)



SETS, a technique inaugurated by Auralog

The software keeps a record of every attempt the learner makes; the learner can therefore go back, listen to them and compare them. In addition, the system allows you to change the threshold of acceptance of the recognition software, so that you can tailor the software to the needs of different students. In addition, with the visual feedback in the form of graphs, Te*L*L me More and Talk to Me also provide the user with feedback in the form of phonetic animations in 3D. These allow the learner to "visualise the articulation of words with dynamic animations and advice on how to formulate words". (http://www.auralog.com.en/tellmemore.html)



The "**3-D phonetic animation**" designed to provide formative feedback to the learner

An additional feature that is worth mentioning is the Internet facility. This facility allows learners to explore pre-selected websites of linguistic interest. Learners can also join Auralog's on-line discussion forums and subscribe to additional services such as a personal tutor.

Te*L*L me More For Kids is aimed at children aged 5 to 11. It is entirely based on speech recognition. It is highly interactive. It is full of rich teaching material and lots of educational games including puzzles, word squares, dictations, loto, spot the difference puzzles and even karaoke.

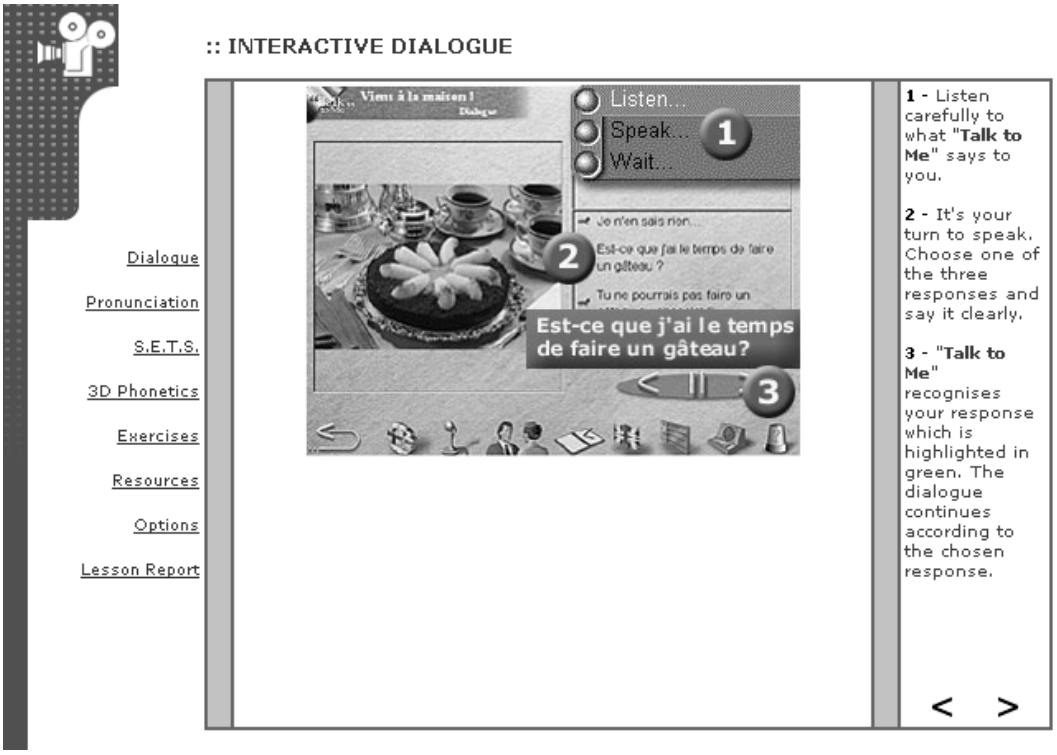
Dyned (http://www.dyned.com)

DynEd produces a range of CALL courseware for learners of English, French, Spanish, Italian, Portuguese, Japanese, Korean, Mandarin and Thai. The good news is that this software is not only available for PC users but also for Mac users (for some languages).

Like Te*L*L me More, the courseware is available in 4 different levels of increasing difficulty. However, much less material is provided with each level (only 30-50 hours of study). In addition, the applications of speech recognition are much more restricted. DynEd does not provide the same level of conversation practice as Talk to Me from Auralog. The nearest it gets to simulating real-life conversations is **branching dialogues**, which have become the most established technique for using speech recognition in CALL.

How does a Branching Dialogue work:

The machine produces a **stimulus sentence** for the learner. It displays it at the top of the screen, for instance "How are you?", it then shows 3 possible answers to this stimulus (3 possible **responses** which, in some cases, can all be correct). The learner is asked to read/produce one of the 3. The machine then recognises one of the 3 (hopefully the one pronounced by the learner) or goes into **remedial mode** to ask the learner to repeat the response sentence.



Branching Dialogue found in **Auralog** "**Talk to Me**". Branching dialogues were first introduced in adventure games.

Examples of other types of exercise that DynEd provides are:

- question-response exercises
- gap-filling exercises
- sentence transformation exercises
-

In brief, with the exception of branching dialogues, speech recognition is used to enhance exercises which would have previously been designed for use with the mouse or keyboard.

Through some of its exercises, DynEd aims to focus on meaning.

"'Click and drag" activities illustrate logical relationships and provide the learner with meaningful practice." (www.dyned.com)

Screenshot taken from Dyned web site

"Combined with Speech Recognition exercises, the **Video Interactions** review key grammar points and introduce useful situational language."

Another of the major manufacturers of CALL courseware that integrates speech recognition is Syracuse Language. Among other products, Syracuse (www.syrlang.com) produces Language for Dummies, Learn Your Way, Smart Start Deluxe, Kids! And TriplePlay Plus!

Language for Dummies, Learn Your Way, Smart Start Deluxe and Kids! are presented at the following address: http://www.syrlang.com/products.html

Here we restrict our presentation to two of their products, namely Language for Dummies and TriplePlay Plus!

Language for Dummies

Priced at $19.99, Language for Dummies can run on any system that has Windows installed. It is available for Spanish, French and English. It aims to develop the four language skills (reading, writing, speaking and listening) through vocabulary exercises, live-action video conversations, interactive games and pronunciation feedback.

Screenshot taken from Syracuse web site. Take part in **real-life conversations**

TriplePlay Plus!

TriplePlay Plus! is available in the US for $49.95, a site license for 5 users costs $250.00, a license for 10 users costs $450. It will run on any machine that has Windows installed and is available for the following languages: English, Spanish, French, German, Hebrew, Japanese and Italian.

Designed for learners aged 8 to adult, the courseware aims to develop three of the four core language skills, listening, speaking and reading.

The main features of the automatic speech recognition component of the system are:

- The SoundStart which allows learners to practice the pronunciation of 50 words using speech recognition and record/playback. (Mackey and Choi, 1998) the record/playback feature allows them to compare their pronunciation with examples provided by native speakers.
- The Letters and Sound, which demonstrates through the use of examples how English letters correspond with the various sounds of the language. (Mackey and Choi, 1998)
- The Language Connect, which gives them access to language learning resources on the Internet. (Mackey and Choi, 1998)

Material is organised in subjects which include food, numbers, home and office, people and so on and so forth. To practice the material learnt in each subject area, the learner can play a number of different types of game. These games are organised in order of increasing level of difficulty. Level one games, which include concentration and bingo, concentrate on the acquisition of individual words. In level two, the vocabulary learnt is used to construct phrases and sentences. In level three, learners use the knowledge they have acquired in levels one and two to take part in conversations that are presented in a comic strip format.

The **Learn to Speak** range from **The Learning Company** (TLC) (www.learningco.com)

In brief this system runs on any machine that has Windows installed on it. The course aims to teach the following language skills, pronunciation, vocabulary, listening, speaking, reading, writing, and grammar through the use of advanced speech recognition, voice recording and playback, talking dictionaries, cultural movies and specialized Internet lessons.

Traci Talk The Mystery (Courseware Publishing International Inc., CA)

“In a series of loosely connected scenarios, the system engages students in solving a mystery. Prior to each scenario, students are given a task (eg. Eliciting a certain type of

information), and they accomplish this task by verbally interacting with characters on the screen. Each voice interaction offers several possible responses, and each spoken response moves the conversation in a slightly different direction. There are many paths through each scenario, and not every path yields the desired information. This motivates the students to return to the beginning of the scene and try out a different interrogation strategy." (Ehsani and Knodt, 1998)

Screenshot from Traci Talk The Mystery
(http://www.clarity.com.hk/English/Software/Titles/ScreenShots/traci.htm)

1.2.6 Advice to teachers (dos and don'ts)

- "Systems should use authentic material." (Aist, 1999)
- "If the goal is acquisition of a skill such as reading or writing, the assisted task should look very much like the unassisted task to maximize transfer to the unassisted task." (Aist, 1999)[1]
- When purchasing software look for a package that comes with a microphone or headset. Headsets are better than microphones. – "Most recognisers perform best when used with a noise cancelling head-mounted microphone. Not only do these microphones filter out extraneous noise, but the head-mounted position ensures that the distance between the speaker's mouth and the microphone is kept at a constant and the amplitude remains stable throughout the utterances." (Ehsani and Knodt, 1998) (if possible purchase software which comes with its own headset as this will guarantee recognition accuracy - it is better to use a headset than a microphone as this ensures that the learner maintains a constant distance from the microphone during use).
- It may be possible to re-use off-the-shelf courseware aimed at native speakers of the target language. The use of dictation software has already been discussed here. It is also worth considering using speech-enabled games written for native speakers of the target language, but these are rare as ASR has hardly penetrated

[1] These two comments apply equally well to any CALL activity.

the game interface. "Obviously, off-the-shelf software does not have all of the CALL specific features desired. However, such software may offer significant price advantages – not to mention glitz – because of its development for the mass market." (Aist, 1999)

1.2.7 Limitations on the use of automatic speech recognition

- speech recognition systems currently exist for only a few of the thousands of languages spoken on the planet. Creating databases for additional languages is currently a costly and time consuming process. However, we should not be discouraged as "techniques for extending the number of languages are becoming more widely understood." (Egan and LaRocca, 2000)
- even more crucially, very few recognisers are trained on non native speech. A recogniser which took a contrastive approach and was trained on non native speech as well as native speech could allow much more accurate feedback to the foreign learner in CALL courseware.
- Just as a human must be able to understand a text to write a dictation, so too must a computer. To address this problem most current systems include some form of linguistic analysis, which is more often than not restricted to syntactic analysis Syntactic analysis does play a role in understanding a text, but there is much more to understanding than linguistic analysis.
- The types of activity that speech recognition enables us to offer language learners are not yet authentic enough. In an ideal world, with the current emphasis on communicative tasks, we would like to offer learners the opportunity to engage in authentic conversations outside the classroom. It has been demonstrated that near recognition accuracy is probably not necessary for simple applications (Ehsani and Knodt, 1998), however to truly emulate real-life conversations, systems must not only be able to recognise speech but to understand it and interact with the user.

1.2.8 Authoring materials with ASR

"Unfortunately not al programs include ASR, and when ASR is available, it is not in the foreign language needed." (Egan and LaRocca, 2000)

Buiding speech recognition courseware for one's student is still not a possibility for the average language teacher as ASR toolkits can be complex to program in, expensive to buy and license and technically difficult to integrate with language authoring programs. Two standard authoring packages available for CALL have allowed this integration, namely **Wincalis** (http://www.humancomp.org/wincalis.htm) and **GLAS** (http://www.blueglas.com/products1.htm). Initially, they integrated with the **Entropic** (http://htk.eng.cam.ac.uk/) toolkit now owned by Microsoft but recently re-released in the public domain.

1.2.9 The future of speech recognition

Future CALL systems which integrate automatic speech recognition may be able to improve both accuracy and feedback by integrating gesture recognition, in particular software that can recognise facial movements. Automatic dictation software that incorporates gesture recognition is already available on the market (for example NLips form Waibel). Gesture recognition dramatically increases recognition accuracy in noisy environments (Graham-Rowe, 1999). It is even possible that this technology could be effectively used in CALL courseware to diagnose errors, as the source of a pronunciation error may be evident from the learners' facial movements (for instance lip rounding).

Assignment 3: ASR in CALL

1. What is the most commonly used method for using ASR in CALL?
2. Visit the Auralog.com and the Dyned.com web sites and compare and contrast the two companies' techniques for using ASR in CALL.
3. Imagine the difference between [ty] in French meaning "you (familiar form)" and [tu] meaning "all", from what you have read above, would you say that a speech recogniser is ready to coach learners to pronounce the difference accurately?
4. What is currently missing in speech recognition driven CALL courseware?
5. Does speech recognition courseware exist for lesser spoken languages such as Irish or Urdu, Swahili, etc... and why?
6. Can a speech recogniser score a learner and what advantage would that have?
7. Does a recogniser need to be aware of the characteristics of non native speech?
8. Give examples of how voice recognition dictation software could be used in language learning.
9. Can ASR help with reading? If so, give examples.
10. Give an example of a visual feedback representation which could be useful in conjunction with the use of speech recognition?

Please note that if you cannot answer more than 50% of these questions, you should read the section above again and make notes, then answer these questions again.

Answers:

1. Branching dialogues used to simulate conversation.
2. Auralog was the first company to use ASR in CALL, they have developed different techniques ranging from the already established "branching dialogue" method used for conversation simulation to more advanced and unique technologies such as their Spoken Error Tracking System. Auralog also uses pitch contours and 3-D phonetic animations as well as more conventional waveform displays. Dyned does has unique methods for the integration of ASR in CALL, they put more emphasis perhaps on remedial CALL by utilising spoken gap-fillers or spoken Multiple Choice Answers. Dyned appears to be the only company capable of using ASR on the Apple Macintosh.
3. What is required to teach the difference between close phonemes such as [y] and [u] is a very high quality recogniser capable of accurately distinguishing the two speech sounds coupled with visual displays/phonetic animations which could show the position of

articulators for each phoneme and also for the phoneme [i] which is a good starting point for the pronunciation of [y]. Indeed being able to show lip movement would enable the learner to easily recognise the distinction between spread lips [i] as against rounded lips [y].
4. Currently speech recognition driven courseware is missing good quality feedback to assist the learner to improve his/her pronunciation and fluency.
5. No, ASR toolkits currently exist only for the main languages of the world, building recognition databases is an expensive business not yet attempted for the lesser known/lesser spoken languages of the world.
6. Yes, some speech recognisers allow thresholds of recognition that can be utilised for learner scoring, Auralog for instance allows scoring from 1 (for very poor non native pronunciation) to 7 (advanced native pronunciation).
7. Very few speech recognisers are trained on non native speech but it would be very useful if they were as they would be able to assist non native speakers much more.
8. Voice Recognition software is typically used for dictation by natives. It can be used by non natives to test their accuracy in dictating in the target language. The main difference with VR is that it is speaker dependent which makes it less usable for foreign language learning.
9. Yes, speech recognition has been used for reading improvements, notably the CSLU toolkit and in work carried out at Carnegie Mellon University (the CMU Listen project, for instance).
10. Visual feedback representation is very useful for pronunciation teaching, for instance, the Syracuse company used a Vowel Interactive Chart to train for vowel accuracy in a product called "Accent Coach".

2. Speech synthesis

The following section represents 5 hours of work including the assignment provided.

2.1 What is speech synthesis?

In lay man's terms speech synthesis is the process of generating artificial speech from a computer. In general, two types of systems exist, those that produce artificial speech from scratch and those that modify naturally produced speech.

Systems which produce artificial speech from scratch can generally be classified into two categories depending on the initial input to the system, these are:

- Concept-to-speech systems
- Text-to-speech systems

Concept-to-speech systems are naturally more complex than text-to-speech systems, because the system must not only reproduce natural articulation, but also reproduce sentence generation.

We will restrict our presentation of how the technology works to the simpler case of text-to-speech systems.

2.1.1 What is speech synthesis used for?

Before going on to look at how these systems work let us consider some real world applications of speech synthesis.

Concept-to-speech and text-to-speech synthesis systems are already being used in a number of applications. The use of speech synthesis is particularly popular in the telecommunications services. Services which make use of speech synthesis include information services such as directory enquiries, and integrated messaging services which allow you to have your e-mail or facsimiles read to you over the telephone. The technology has also already been exploited to produce talking books and toys.

Speech synthesis systems could also be used to provide aids to handicapped people, in particular if they are coupled with an Optical Character Recognition (OCR) system, they can be used to give blind or visually impaired people access to written material.
As speech is the most natural form of communication, speech synthesis can be integrated into machines to facilitate man-machine communication. In particular it can be used to provide a vocal interface for blind or visually impaired people. (Dutoît, 1997)

The modification of naturally produced utterances or waveform editing has mainly been used for creating sound effects.

2.1.2 What is a waveform editor and how does it work?

Waveform editors are used to modify naturally occurring speech. These systems take an utterance produced by a human and display its waveform (spectrogram). The user can then edit the waveform, for example they can increase its amplitude and frequency. Having done this the computer will re-synthesise the original utterance to take account for the changes that have been made. This process is known as **re-synthesis**.

2.1.3 Text-to-speech systems – how do they work?

Before continuing to describe how a text-to-speech system works, it is necessary to clarify the meaning of text-to-speech synthesis. Two meanings of text-to-speech synthesis are in general use. For some, text-to-speech synthesis covers the whole process from the input of text to the production of an acoustic signal. For others, text-to-speech synthesis only refers to the first stage of the full process, that is the production of a narrow phonetic transcription of the text. Here we will present the entire process.

2.1.4 Global architecture of speech synthesisers

In general, text-to-speech synthesis can be broken down into two distinct tasks:

- The production of a narrow phonetic transcription of the text, that is to say one representing the allophones (the realisation of different phonemes in context) of the utterance.

- The conversion of this phonetic transcription into acoustic signals.

Most systems are therefore composed of a **Natural Language Processing** (NLP) module, responsible for producing a narrow phonetic transcription of the text, and a coder or **Digital Signal Processing** (DSP) module, responsible for converting the transcription into an acoustic signal.

2.1.5 The NLP module

In principle, there are two stages to the production of a narrow phonetic transcription of a text.

(1) **Phonetisation or letter-to-sound conversion**, which is the creation of a broad phonetic transcription from an orthographic text.
(2) **Prosody generation**, which is the generation of the melodic contour of an utterance and the duration of each phoneme. This stage is essential for producing natural sounding synthetic speech. (Goldman, Laezlinger and Wehrli, 1999)

In order to phonetise a text, systems can either use an exhaustive lexicon (vocabulary) containing a phonetic transcription of every word that the system is likely to encounter or a rule-based system which converts letters or groups of letters to a transcription of their pronunciation.
A rule-based letter-to-sound conversion system is the minimal requirement for the NLP module of a text-to-speech system. It is also an essential component of lexicon-based systems because no lexicon can ever be entirely exhaustive. In a lexicon-based system, the letter-to-sound conversion system permits the phonetisation of out-of-vocabulary words, typically these are neologisms and proper and foreign nouns.
In order to treat problems such as heterophonic homographs, words that have one orthographic form but several phonetic realisations (for example ‘desert’ in “She wished she could desert him in the desert” [The Homograph Page]), and to determine the prosody of an utterance, most speech synthesis systems will have a much more sophisticated NLP module.
The most important component of such systems will be a parser or syntactic analyser, whose prime objective will be to determine the grammatical categories of the words which make up the sentences and to establish the relationships which hold between them, in other words determine the syntactic structure of the sentence.

2.1.6 The Digital Signal Processing (DSP) Module

Finally a coder converts the narrow phonetic transcription into an acoustic signal.
Several methods of signal processing have been used in the history of speech synthesis.

The two most commonly used methods are:
- parametric coding
- concatenation

Parametric coding aims to electronicaly reproduce natural speech, by simulating the different parameters (or acoustic features) of the speech signal. The most significant developments in this domain are **formant synthesisers**. Formant synthesisers are based on the assumption that it is possible to satisfactorily model speech by electronically reproducing only those features which are necessary from the point of view of perception, that is the frequency and amplitude of the formants. (Styger and Keller, 1994)

"Concatenative synthesis is based on signal processing of natural speech databases." (TTS – State-of-the-art) This technique aims to reproduce utterances by putting pre-recorded segments of human speech together end-to-end.

2.1.7 Parametric coding vs. concatenation

Two main factors distinguish these two methods. Firstly, concatenation is based on naturally produced speech, whereas parametric coding produces sounds which are entirely synthetic or artificial. The second difference between the two techniques is the way in which they control the different speech parameters. In concatenative speech synthesis, coarticulation and other articulatory affects can be modelled automatically through the choice of segment. With parametric coding, each of the parameters involved in producing these effects must be modeled through the use of rules.

Before going on to present the products that are currently available on the market we should first consider the characteristics that distinguish one system from another. These are the factors which should be taken into consideration when purchasing a speech synthesis system. These are the very same features that distinguish one speech recognition system from another, the only difference is that with speech recognition these factors apply to the speech input, with speech synthesis they apply to the output. These factors are:

- mode/type of speech output
- degree of dependence on the speaker (speaker dependence)
- size of the vocabulary

2.1.8 Mode of speech output

In principle, there is a dichotomy between systems which produce discrete speech, that is with pauses between the individual words of an utterance, and those which produce continuous speech output.

2.1.9 Speaker dependence

Traditionally, a distinction is made between speaker dependent and speaker independent systems. Speaker dependent systems only produce speech in **one voice**, whereas speaker independent systems can produce speech in a number of **different voices**.

2.1.10 Vocabulary size

Intuitively, we would be tempted to conclude that the larger the vocabulary, the more accurate the speech produced. However, due to the creative nature of language, the lexicon (vocabulary) of a speech synthesiser is never exhaustive; it can never include all the words of a language. There will always be a need for a letter-to-sound conversion module to process neologisms, derivations and proper nouns. In fact what is of real importance is not the size of the vocabulary but the choice of vocabulary items. In order to save on storage space, the vocabulary of a speech synthesiser should only contain words whose pronunciation (phonetic transcription) cannot be accurately predicted through the application of letter-to-sound conversion rules. In other words, an economical lexicon will only contain words and derivations with irregular pronunciation and the roots or canonical forms of derivations.

2.1.11 Generic products

The "Museum of Speech Synthesis in Grenoble" has links to some speech synthesis systems which you can test for yourself. It can be found at the following URL: http://www.icp.inpg.fr/ICP_old/equipes/synthese/musee.en.html

Three of the major examples of commercial speech synthesis software are produced by:

- IBM
- AT&T Bell Laboratories (Lucent)
- Lernout & Hauspie (see note higher up on L&H)

2.1.12 IBM

IBM ViaVoice Text-to-Speech produces continuous speech in the following languages: US English, UK English, French, Italian, Spanish, German, Brazilian Portuguese, Japanese and Chinese in an unlimited number of different voices. The software is based on formant synthesis, consequently it is easy to modify speech parameters to reflect differences in gender, pitch, head size, roughness and so on and so forth. The software can be run on a number of different platforms including Windows, AIX, LINUX, and Solaris. When used in a Windows environment, it is particularly powerful/useful because it can be used to speech enable any application. It can be run on the desktop, over a telephone network, on the Web or in a car. Users can therefore access information anytime, anywhere and with almost any device. Some examples of how it can be used are:

- to read through word processed documents
- to listen to e-mails from a wireless telephone
- to surf the Web while sitting in traffic
- to use an in-flight phone to get stack quotes

The software is available as a Speech Development Kit (SDK) (IBM ViaVoice TTS SDK for Windows) and in a run time version (IBM ViaVoice TTS Run Time Kit for Windows). (ViaVoice text-to-speech: the voice of IBM).

2.1.13 AT&T Bell Laboratories (Lucent)

AT&T Bell Laboratories produce a range of speech synthesis products based on the technique of concatenation. These systems, which can be run on Windows or UNIX machines, currently support the following languages: English, Chinese, German, French, Italian, Latin American Spanish, Romanian and Russian. Some are available in both a female and a male voice. The laboratory is continually working on techniques to increase the naturalness of the speech produced and expanding the set of languages supported. The Bell Labs Text-to-Speech system has various application including reading electronic mails messages, generating spoken prompts in response systems, and as an interface to an order-verification system for salespeople in the field. (Bell Labs TTS project).
A simple description of the system architecture is available along with links to examples of the speech produced by the system can be found at the following Web page:
http://www1.bell-labs.com/project/tts/

2.1.14 Lernout & Hauspie (L&H - see note higher up on this company now owned by Scansoft)

To give a slightly different perspective on speech synthesis, we have chosen to present Lernout & Hauspie's Kurzweil speech synthesis software. This is an example of speech synthesis software specially designed for use by blind or visually impaired people. It uses concatenative speech synthesis to produce continuous speech. The speech synthesis program is coupled with a scanner and Optical Character Recognition (OCR) software, to enable users to read any text. The speech rate can be varied from 200-600 words per minute. Additional features include the ability to send the scanned files to portable note taking devices and to create MP3 audio recordings of the scanned documents. Unfortunately the Web site does not indicate which languages Kurzweil supports. Lernout & Hauspie's general speech synthesis products can be run on any machine that has Windows installed, and can speak the following languages in both a male and a female voice: English, German, Dutch, French, Spanish, Italian and Korean. Unfortunately, due to the transfer of ownership of L&H, details of the Kurzweil 1000 reading machine can only be found on the inventor's original URL:
http://www.kurzweiledu.com/k1000readit.html

Links to the producers of other speech synthesis products and a summary of the main features of these products can be found on the following Web page:
http://www.acoustics.hut.fi/~slemmett/dippa/appb.html

Assignment 4: Defining Speech Synthesis

1. This module has restricted its consideration to one broad type of speech synthesis, what is it called?
2. What are the two main types of digital signal processing used in the final part of synthesis and what is the difference between the two techniques?
3. Visit the ICP Museum of Speech Synthesis web site, listen to the same French sentence spoken by 4 different synthesizers, rate each one for its natural sound quality (closest to human speech).
4. What is meant by re-synthesis?
5. Give examples of popular uses of speech synthesis in the general communication arena.
6. Describe the two elements of the Natural Language Processing part of a speech synthesizer.
7. By what parameters is synthesis described and compared (as is speech recognition)?
8. Is speech synthesis used in the assistive interface (blind people, etc...)?
9. Listen to the examples of synthesis on the Bell Labs Lucent Technologies web site in the language of your choice. Describe how you could use the various "voice" options in language learning.
10. Before studying the actual use of Speech Synthesis in CALL, list in bullet points the uses which you could imagine making in your own pedagogy.

Please note that if you cannot answer more than 50% of the questions above, you should read the section again making notes and then make a new attempt to answer these questions.

Answers:

1. Text to Speech is the common name of the most popular type of speech synthesis.
2. Digital Speech Processing can be achieved via parametric coding or concatenation. Parametric coding produces completely artificial speech via the use of the main parameters in speech. Concatenative synthesis relies on putting end to end segments of naturally produced speech that can take more account of co-articulation phenomena, as a result, concatenation produces more natural sounding synthesis.
3. My least favourite synthesizer is the KTH which has a "foreign" accent, next is the ICP, the CNET synthesizer provides yet more improvement but the most natural sounding for me is the most recent by LAIP in Lausanne.
4. Re-synthesis is the process by which a speech sequence is re-processed after modifying its parameters, this can be done to synthesized speech or real speech.
5. Telecommunications make heavy use of speech synthesis, it is present in telephone systems for instance. Synthesis is also used in other devices such as lifts or toys, etc...
6. The two main stages of Natural Language Processing in a synthesizer are letter to sound phonetisation (affecting individual phonemes) followed by prosody generation (rhythm and intonation) affecting the longer speech segment.
7. Broadly, speech synthesis is described and rated by similar parameters as speech recognition, these include mode of speech output (discrete versus continuous), speaker

dependence (one voice) as against speaker independence (multiple voices) and vocabulary size (small, limited application tools as against large vocabulary systems).
8. Yes, speech synthesis was used very early in reading text for blind people (see Kurzweil reading machine used in particular by Stevie Wonder).
9. One obvious use of speech synthesis in CALL could be the slowing down of speech to make it easier to understand, this is illustrated on the Bell Labs site which also features male, female and kid voices.
10. Speech synthesis could be used in a variety of ways in CALL, for instance to slow speech down, to produce speech without having to record native samples, to re-synthesize the speech of learners and show how modifying parameters affect their pronunciation. It could also be used as a mechanism for speaking the text produced by a non native learner to hear how the piece of text produced sounds.

2.2 Speech synthesis in CALL: the following, longer section represents 9 hours of work including the assignment provided.

Although speech synthesisers are more reliable than speech recognition systems their use in language learning has been rather limited (Dutoît, 1997). Currently there are no commercially available products which integrate speech synthesis.

We will therefore consider the potential advantages of integrating speech synthesis into language learning applications followed by some examples of experimental projects.

The integration of speech technology in language learning applications has many advantages. It is in fact essential because the study of phonetics can be an integral part of advanced language learning.

The first advantage of integrating speech technology into CALL software is that it allows us to make the transition from passive to interactive learning. This is very important, because experiments have demonstrated that we remember 10% of what we read, 20% of what we hear (listen to), 30% of what we see, 50% of what we see and hear, 80% of what we say and 90% of what we say and do at the same time. (Brierley and Kemble (eds), 1991).

One of the major advantages of the integration of speech synthesis in CALL applications is that the tool can be used as an inexhaustible/indefatigable tutor: learners can listen to the examples as many times as they want.

In addition, both concept-to-speech, text-to-speech and waveform editing permit the creation of types of examples that a human is physically incapable of producing (for example, utterances with intonation but no rhythm). Such examples may be useful in the context of language learning if the teacher wishes to isolate and emphasise a specific feature of the target language (Keller and Zellner-Keller, 2000). It is easier to teach one notion at any one time, so the isolation of parameters can play a very useful role in spoken language pedagogy.

2.2.1 Speech synthesis in language learning – state of the art

There are far fewer researchers working in this area than in the domain of speech recognition in CALL. Skrelin and Volskaja are one of the exceptions in an article "which outlines the use of speech synthesis in language learning and lists dictation, distinction of homographs , a sound dictionary and pronunciation drills as possible applications." (News from CTICML)

2.2.2 The use of concept-to-speech and text-to-speech systems

As we have previously stated, the use of speech synthesis in language learning is a relatively new idea, most research has not yet found its way out of the research laboratory. There must, however, be significant uses of synthesis in spoken CALL.

Myers (2000) describes the use of hand-held electronic dictionaries with speech synthesis facilities by Chinese learners of English. In particular, she reports on a study into their use for reading comprehension. The study demonstrates that there are two major advantages of using dictionaries which not only provide the orthographic form of a word along with a definition but also its pronunciation. Firstly, they help consolidate vocabulary. Through being able to repetitively both see and hear target language words, learners eventually develop the ability to tell whether a word sounds or looks right just as a native speaker does. Secondly, this technology has proved useful in silent reading exercises. Learners often find that they actually do know a word which they did not recognise in orthographic form, once they hear it pronounced.

Another project which aims to integrate speech synthesis into a CALL environment is **FreeText** (French in context: An advanced hypermedia Computer-Assisted Language Learning (CALL) system featuring Natural Language Processing (NLP) tools for a smart treatment of authentic documents and free production exercises.) This project is still in the developmental stage. Eventually, it is hoped that the whole interface of the FreeText programme will be voice-enabled, that the learner will be able to use the speech synthesis system to listen to the mentors who will be a guide to the tutorials, to listen to the authentic documents around which the tutorials are centred, to listen to explanations of grammatical points, instructions to exercises, the exercises themselves and the answers to those exercises. Through the use of speech synthesis, the project aims to go one step further in making language learning truly interactive. The integration of speech synthesis will give the learner additional listening practice and will also help consolidate vocabulary and grammar.

The FreeText project will integrate FIPSVOX, the speech synthesis system developed at the University of Geneva. It is hypothesised that the output of FIPSVOX will be much more reliable than that of most other speech synthesisers, as unlike most synthesisers on the market, it is based on true linguistic processing rather than on statistical models, and will therefore be suitable for learners.

A fuller description of the Freetext project is available at the following address:
http://www.latl.unige.ch/freetext/en/description.html

You can also try out the synthesiser FIPSVOX yourself by following the link “synthèse de la parole” from the following web page: http://www.latl.unige.ch

A discussion of the potential use of speech synthesis in CALL is made in an article by Bob Godwyn Jones published by the Language Learning & Technology web-based journal. The full article can be read at http://llt.msu.edu/vol3num2/emerging/#3

Godwyn-Jones also mentions the new area of Speech to Speech processing which combines speech recognition and speech synthesis to produce speech translations from one language to another. Work in this area was pioneered by CMU in the 80s and continued by Lernout & Hauspie who had ambitions to release portable speech translators of this kind.

A novel use of speech synthesis in CALL is found in a prototype game constructed by Delcloque, Toche and Métivier in 2001. This science fiction game is called "Zoé va au Zoo, Odyssée Spatiale de l'An 01", in the game, the heroine Zoé is assisted by a fairy, an animated character which comes with the L&H Realspeak female synthetic voice. The resource was chosen because it procured the most natural sounding synthesis on the market. The synthetic voice also has the advantage of speaking in an authentic yet pedagogical manner which makes it intelligible to the target audience of young learners (9 to 11).

2.2.3 The use of waveform editors

According to Bonneau, Laprie and Colotte (2000), the fundamental idea behind the use of waveform editors in CALL is to allow the teacher to draw the learner’s attention to certain aspects of speech, in particular intonation and the pronunciation of certain sounds. They discuss the use WinSnoori, a waveform editor which allows you to manipulate both the speed and fundamental frequency (tone) of utterances. The following is a brief summary of some of the ways in which WinSnoori can be exploited in language learning.

- It can be used to focus learners’ attention on rises and falls in intonation through the multiplication of fundamental frequency.
- It can be used to emphasise the pronunciation of certain sounds (or segments) such as stops, bursts and fricatives, by enhancing (or augmenting) their spectral pattern.
- Utterances or parts of utterances which are difficult to perceive can be lengthened through the manipulation of the speech rate.

The waveform editor WinSnoori is presented on the following Web page, which has links to a guided tour of the interface and explanations of how WinSnoori can be used: http://www.babeltech.com/winsno/winsno.html

Waveform editors can also be used in conjunction with speech recognition programs to give feedback to the learner. In particular they can be used to derive correctly articulated utterances from learners’ incorrectly articulated utterances. Aist (1999), points us to the

research of Nagano and Ozawa, who demonstrated that using a waveform editor to manipulate learners' own utterances to match the prosodic contours of the target language was more effective than using recordings of native speakers.

Many other examples of waveform editors are presented on the following web page: http://liceu.uab.es/~joaquim/teaching/Phonetics/fon_anal_acus/herram_anal_acus.html

2.2.4 Conclusion

In conclusion, the major limitation of the current speech synthesis systems is that many can only 'speak' in one or two voices and one style, which is more often than not reading style. This is a serious limitation; according to estimates made by Keller and Zellner-Keller (2000), just taking into account variations in speech rate, type of speech (spontaneous, prepared oral, command, etc) and material (continuous text, lists, etc) there are at least 180 different speaking styles. Reproducing these styles and speaker variation is extremely difficult. Reproducing them through parametric coding requires the manipulation of a large number of parameters. Reproducing them through concatenative synthesis requires the creation of a different speech database for each voice and each style of speaking.

The quality of synthesised speech can also still be dramatically improved. Most synthesised speech still sounds 'robotic'. Several factors are responsible for this:

- inaccuracies in simulation
- lack of prosody or incorrect treatment of prosody
-

In addition, partially due to the lack of prosody and also due to the fact that we are quite justified in saying that computers do not know what they are talking about, synthetic speech also lacks emotion.

It is hoped that advances in both Natural Language Processing (NLP) and Artificial Intelligence (AI) will eventually provide solutions to these problems.

Other potentially useful information

Keller and Zellner-Keller (June 2000) suggest that speech synthesis can be exploited in the following activities:

- "A good model of language is particularly useful in the training of prosodic and articulatory competence. A speech synthesiser can slow down stretches of spoken language at will, which eases familiarisation and articulatory training with novel sound sequences."
- "Advanced learners, on the other hand may wish to experiment with accelerated reproduction speeds. Those are commonly used by the visually handicapped for scanning of voluminous text rapidly."
- "Another obvious application is listening comprehension."
- Speech synthesisers can also be used as a "substitute native speaker"

3. Audio-visual synthesis

3.1 What is audio-visual synthesis?

Audio-visual synthesis, or multimodal synthesis, is the generation and synchronisation of facial (and sometimes gesture animation) animation with speech, whether natural or synthetic.

Currently there are a number of speaking Audio-Visual Speech Synthesizers (also referred to as "Talking Heads") working with the following languages: English (University of Carlifornia Santa Cruz, Miralab, HMS, VisLab, DEC, British Telecom), Swedish (KTH), Japanese (Sony and NTT) and French (ICP and INA).

Some examples can be seen or accessed from links on the following pages:

Beskow (2001) Multimodal Speech Synthesis http://www.speech.kth.se/multimodal
Speech perception by ear and eye http://mambo.ucsc.edu/psl/pslfan.html

Speech perception by ear and eye also contains a link to a demonstration of the McGurk effect, an important issue in "visual speech perception".

3.2. Why use audio-visual synthesis?

Visible speech, which means the speech which can be read from facial and lip movements, dramatically increases intelligibility in situations where auditory speech is degraded – for example, in noisy environments, when using applications with limited bandwidth, and when people have hearing impairments. “If, for example, only roughly half of a degraded auditory message is understood, its pairing with visual speech can allow comprehension to be almost perfect.” (Massaro and Cole, 2000) Visual speech is extremely successful because:

- Visual speech is highly robust: You can lip-read when not looking directly at the speaker’s face and lip-reading remains accurate even when the face is blurred. (Massaro and Cole, 2000)
- Visual speech provides complementary information to auditory speech:

> “For example, the difference between, /da/ and /ba/ is easy to see, but relatively difficult to hear. On the other hand, the difference between /ba/ and /pa/ is relatively easy to hear, but very difficult to see.” (Massaro and Cole, 2000)

(Lessard, 1996)

> The difference between /b/ and /d/ is easy to see because /b/ is a bilabial consonant and /d/ is and apico-dental consonant. As far as the difference between /b/ and /p/ is concerned, it is audible but not visible because the only difference between their articulation is that /b/ is voiced (the vocal cords vibrate) and /p/ is voiceless.

- Auditory and visual speech are optimally integrated.

Even when auditory speech is not degraded, there is strong evidence that speakers integrate visual information. "Conflicting auditory and visual information can distort the perception (McGurk effect)" (Paula Web Survey) "For example, if the ambiguous auditory sentence, 'my bab pop me poo brive', is paired with the visible sentence 'My gag kok me koo grive', the perceiver is likely to hear '**my dad taught me to drive**'." (Massaro and Cole, 2000).

Speech is both verbal and visual. "Visual signals are used in conversation to signal feedback and turntaking." (Beskow et al, 2000) In addition to conversational signals, visual signals also function as linguistic signals; "eyebrow movements and nodding for accentuation can function as parallel signals to intonation", "blinking, changes of gaze, eyebrow raising, frowning, head nodding, and head turning" (Beskow et ali, 2001) may be used to emphasize stress placement and phrasing.

- It adds realism.

3.3 What is audio-visual synthesis used for?

Current applications of audio-visual synthesis include:

- research on human communication and perception
- tools for people with hearing impairments
- multimodal agent-based user interfaces (for example, "this technology is well-suited for information systems in public and noisy areas, such as airports, train stations and shopping centres." (Beskow, 1996)

3.3.1 Why use audio-visual speech in language learning?

The reasons for using audio-visual speech in language learning are much the same as those for using it in other situations.
The following features of audio-visual synthesis systems are of particular interest to language learning:

- Like speech synthesis systems, audio-visual synthesis systems can act as an indefatigable substitute native speaker.
- many audio-visual synthesis systems can demonstrate articulation. Articulator movements normally hidden from the learner's view can be seen, by rendering the skin of the model transparent or by presenting sagittal sections. (Beskow et al, 2000) Being able to see articulator movements improves perception and reinforces the relationship between spelling, sounds, and articulations.
- In addition, articulations can often be emphasized through super or hyperarticulation.
- "Subjects listening to a foreign language often incorporate visual information to a greater extent than do subject listening to their own language." (Beskow et al, 2000)
- Visual cues such as nodding and smiling can be used to encourage the learner.

3.4 How are animated faces created?

The methods used in facial animation are essentially the same as those in speech synthesis. Facial animation can be generated electronically through parametric coding, or it can be generated by concatenating pre-recorded images from natural speech databases.

Systems based on parametric coding may or may not be anatomically motivated. Those which are not anatomically motivated are generally based on observation. In general they are based on the manipulation of a wire frame model the surface of which is texture mapped to provide the effect of skin. Anatomically based synthesizers attempt to simulate muscle movements.

Muscle-based models are more attractive as they are more elegant, however they are much more difficult to program. Both concatenative and wire frame models are much simpler. The advantage of wire frame models over concatenative models is that they require significantly less data storage. In addition, wireframe models are much more flexible than concatenative models, because they allow you "to control facial features independently of each other." (Beskow, 1996)

3.5 Audio-visual synthesis: the state of the art

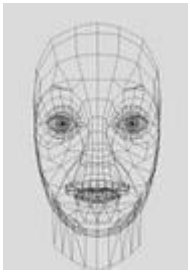
The two major centres for research on multi-modal synthesis are: KTH Stockholm and the Centre for Speech Language Understanding in Oregon, USA (CSLU).

3.5.1 KTH Stockholm

The audio-visual speech synthesis system implemented at KTH Stock holm is based on the synchronisation of the KTH rule-based formant text-to-speech synthesizer and the parametric face model developed by Parke (1982)

"The Parke facial model consists of a mesh of about 800 polygons, that approximate the surface of a human face including eyes, eyebrows, lips and teeth. The polygon surface

can be deformed using 50 parameters, that move the vertices of the polygon network in different ways." (Beskow, 1996)

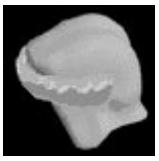


Holger (1995) (Beskow, 2001)

Holger

August (1998) (Beskow, 2001)

At the time of publishing (Beskow, 1996), the audio-visual synthesizer developed at KTH did not allow users to see inside the model, learners could not see the movement of the articulators. The system only modelled tongue movements that could be seen from the outside of the model, that is apical tongue motion. Pictures on the web site Multimodal Speech Synthesis suggest that today the tongue is modelled in full.



Tongue and teeth (Beskow, 2001)

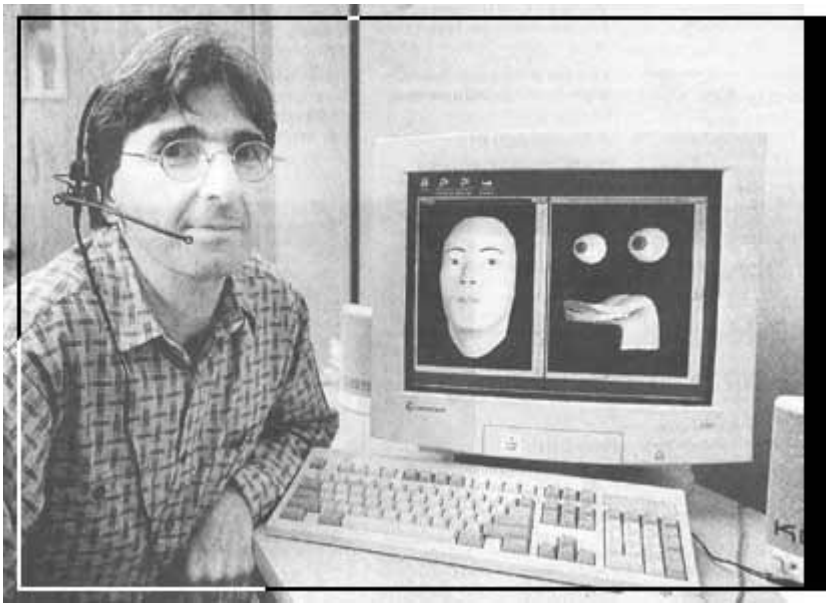
An interesting feature of the system is that it "allows animation not only of the facial model, but also of, for example, general jointed character bodies, as needed for animation of gesturing agents." (Beskow, 1996)

Olga (1996) (Beskow, 2001)

Demonstrations of all the different audio-visual speech projects can be accessed from the following web page: http://www.speech.kth.se/multimodal/video/index.html
This is an invaluable site but you will need a university type (T1) fast network or a broadband connection to view the talking heads videos.

3.5.2 CSLU

Audio-visual speech synthesis software is available free of charge, under a license agreement for educational, research, personal or evaluation purposes, as part of the CSLU toolkit.



Baldi (source http://www.oraldeafed.org/schools/tmos/news-050698.html)

The aim of the CSLU Toolkit project is "to provide researchers with the knowledge and tools to advance the state of the art and powerful tools that enable even inexperienced users to rapidly design, test and deploy spoken language systems" (The CSLU Speech Toolkit) In order to achieve these aims, in addition to facial animation, the toolkit comprises:

- Speech recognition technologies including artificial neural network (ANN) classifiers, hidden Markov models (HMM) and segmental systems.
- Speech synthesis using the Festival text-to-speech system. Speech can be produced in six voices, including male and female versions of American English and Mexican Spanish.
- Authoring tools
- Waveform analysis tools
- Programming environment

3.5.2.i Facial animation

The audio-visual speech character created at CSLU is called Baldi. Baldi is based on the synchronisation of two technologies, text-to-speech synthesis and facial animation. The text-to-speech system used is the Festival system, which was developed at the University of Edinburgh in Scotland. The facial animation program was developed at the University of California Santa Cruz.

The facial animation system is based on parametric coding. Parameters are used to control a wireframe model which is texture mapped with a skin surface.

Particular features of Baldi are:

- Although Baldi was designed to be controlled by text-to-speech synthesis, Baldi's facial animation can also be aligned with natural speech if this is deemed more appropriate.
- In addition to the wire frame model of the exterior of the face Baldi incorporates a 3D model of the tongue.
- Speech is visible not only from the outside of the face, but also from the inside. "The skin of our talking head can be made transparent so that the inside of the vocal track is visible, or we can present a cutaway view of the head along the sagittal plane." (Massaro and Cole, 2000)
- Baldi's head can rotate, so that it is possible to view the head from the back. This is believed to be more conducive to language learning than a frontal view, for "the tongue in this view moves away from and towards the student in the same way as the student's own tongue would move." (Massaro and Cole, 2000)
- Baldi can be used to enhance visual cues through hyperarticulation. For example, "to distinguish /k,g/ from /t,d/, the jaw can be moved downward to a greater extent." (Massaro and Cole, 2000)

3.5.2.ii Using the Toolkit in the classroom

An example quoted in The Oregonian Reports on Baldi:

"To create conversation with Baldi, all the teacher has to do is use programs from the CSLU Toolkit. Teachers can type in the words they want Baldi to say and the words they want Baldi to recognise in response."

The toolkit and authoring program can be downloaded by following the relevant links from the following web page: http://cslr.colorado.edu/toolkit/main.html

Assignment 5: Speech and A-V Synthesis in CALL: Test your knowledge

1. Write a summary of the existing uses which have been made of Audio Speech Synthesis in CALL.
2. Explain the use made of A-V talking heads.
3. Now extrapolate and provide bullet points on how audio or A-V synthesis could be used in CALL.
4. What do you perceive are some of the factors which have stopped teachers and researchers finding ways of using such promising technologies in CALL?
5. Can you see a use for waveform editors and re-synthesis?
6. What are the main ambitions of the Freetext project (check this on their web site too).
7. Write a short scenario of a game used in language learning that could incorporate both speech recognition and speech synthesis, how would you use the said technologies?
8. Do a web search using Google.com and find out whether Speech Synthesis can also be used on an Apple Macintosh, if so what is it called?

9. Explain the idea of Visible Speech, why does it matter?
10. Could A-V synthesis help in creating digital characters capable of expressing anthropomorphic emotions?

Please note that if you are not capable of answering more than 50% of the questions above, you should read this section again and then attempt to answer the questions.

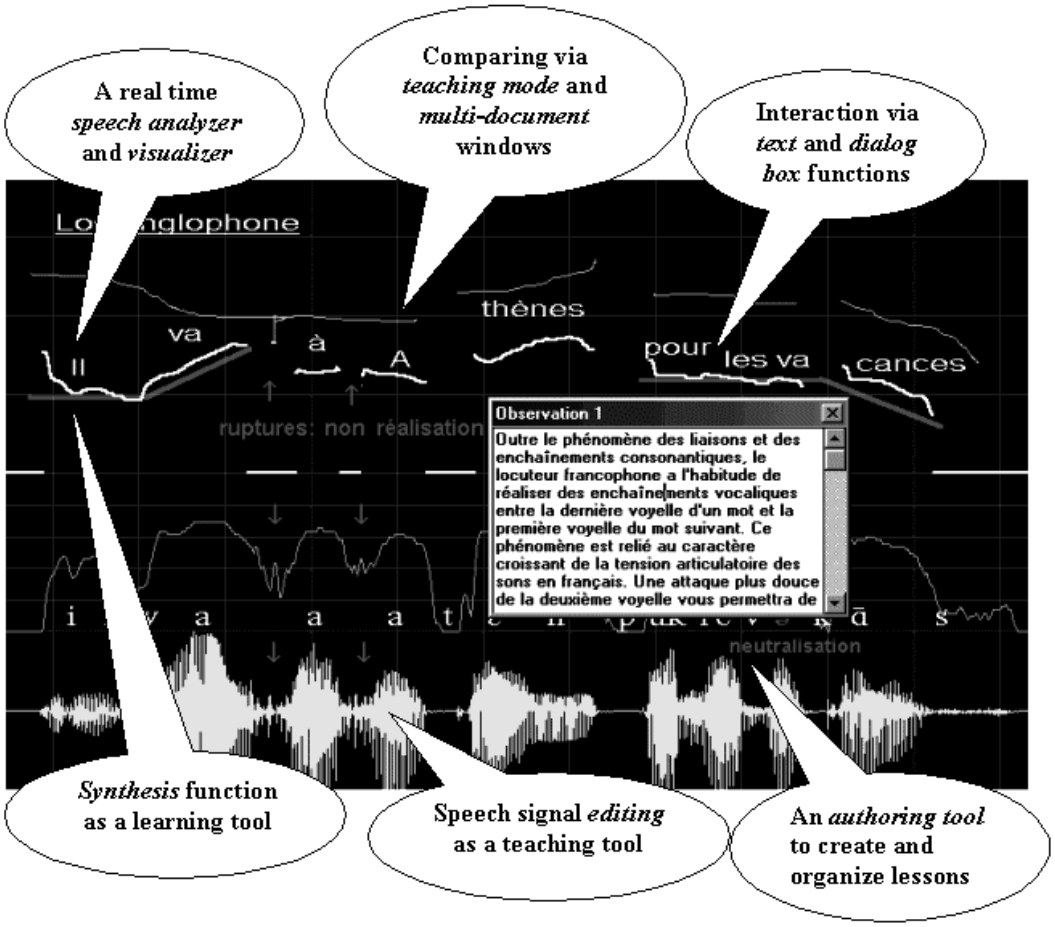
Answers:

1. Audio Speech Synthesis has been used to assist reading, to slow down speech or accelerate it in listening comprehension, to illustrate phonetic differences for individual speech sounds and also for intonation.
2. A-V talking heads have hitherto been used for the re-education of deaf children but they can have many more functions in CALL.
3. One can imagine using Talking Heads and animated agents in a variety of ways, for instamce, a talking head can be used to show how to articulate a sound, it is now possible to show inside and side view of articulators. In time, with progresses in recognition, a talking head could be used to show how a learner is articulating a sound or a sound sequence. Talking heads and animated agents will make interesting characters in games. Once such technologies are available for the mass market of language teachers, their creativity will find many ways of utilising them.
4. The unavailability of speech synthesis for the average language teacher must have been one of the main reasons why synthesis has not been used as heavily in CALL, also, voices need to sound more natural, less robotic. Another problem has been the integration with "ordinary" CALL authoring software.
5. Winsnoori proves that several good uses of waveform editing and re-synthesis may be used, these include a focus on learners' attention to intonation patterns (the intonation may be changed by manipulating the fundamental frequency (F0). Synthesis can be used for emphasizing the pronunciation of certain sounds and segments to bring attention to particularly problematic phonemes. Some sounds can also be lengthened to make them easier to perceive.
6. Freetext aims to provide an integrated environment for language teachers to author materials and use speech synthesis to illustrate phonetic and linguistic points.
7. One can imagine a detective story such as Oscar Lake where the entire interface is driven by sound and speech instead of keyboard touches, some characters could speak with synthetic voices, there could also be digital (pre-recorded) speech, the learner could interact/play by using their voice and the game would progress through the use of speech recognition. This would be a much more naturally engaging interface.
8. Apple Macs were among the first computers to use Text to Speech Synthesis, the initial name was MacinTalk, the technology is now called Plaintalk which includes some limited speech recognition in a few European languages.
9. The notion of visible speech is that speech is perceived by human beings via two main media, sound and movement, facial and lip movements are therefore essential to accurate speech perception. Optimum perception is reached when both senses are at play, Different phonemes are perceived more easily via Visible Speech, others via Auditory Speech.

10. The link between A-V Synthesis and the expression of human emotions is very strong as witnessed by the interest shown in both by scholar Dominic Massaro, one of the main advocates of the importance of visible speech.

4. Visualisation: this last section represents 5 hours of work including the assignment provided.

The final thematic area within STiLL is the promising domain of visualisation. This is an extension from speech analysis and indeed the tools used reflect this. Here we study the best tool available for Windows for presenting visual representation of speech segments. Firstly, we need to explain the two main areas which parrallel the study of phonetics. Visualisation can concentrate on **individual speech sounds**, on showing their formants, on recognising them and utilising this knowledge to improve the pronunciation of the phonemes of the target language. Of even greater importance is the display on screen in real time of **suprasegmental features**, particularly **intonation contours**, this can also help understand **rhythmic** phenomena better. The tool described below was invented by Philippe Martin and has been used very effectively by a number of leading researchers and language teachers, the most prominent perhaps being Aline Germain Rutherford. More information about this tool can be found from the **Winpitch** web site at http://www.winpitch.com/ from which we have extracted the screenshots below:

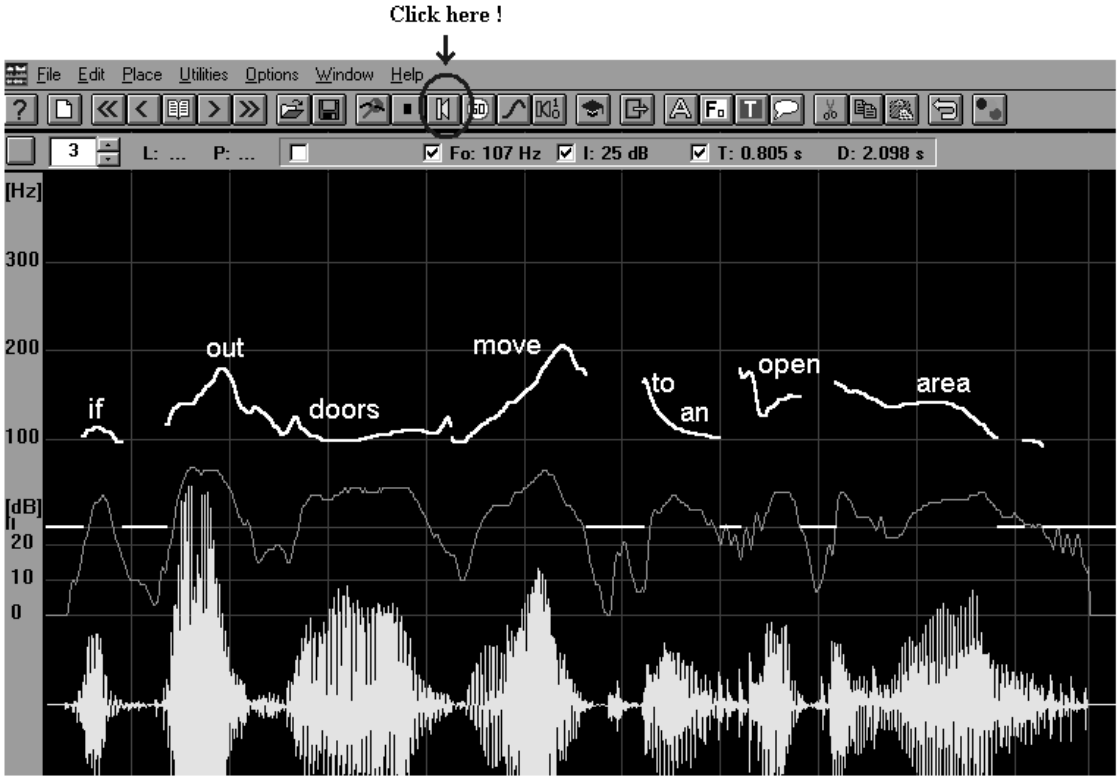


WinPitch LTL is a “user friendly” real time speech **analyser** and **visualiser** which can be used by any language teacher and/or student to facilitate the teaching/learning of any second language. The tool is primarily used for pronunciation training, it includes an authoring tool for teachers and can be used for asynchronous communication. Winpitch files can be sent and received by e-mail creating a pedagogic exchange between learner and teacher.

One of the prime reasons for speech training is what is referred to in the literature as "phonological deafness". This is the idea that when one starts learning a foreign language, one does not hear sound distinctions which are not made in one's own language, this may be the case for instance for the two types of "i" used in English, the higher, longer and tenser [i:] found in the word "seat" and the slightly lower, shorter and laxer [i] found in the word "sit". An explicit approach to pronunciation teaching fostered by the use of a tool such as Winpitch is likely to result in the learner losing this phonological deafness sooner and articulating target sounds accurately more quickly too.

As stated on the Winpitch site "Research shows that **visual feedback** can accelerate and ease the acquisition of sounds, intonations and prosodic patterns and that it is also an important motivational factor".

The following Winpitch screenshot shows 3 different types of graphic representation, the standard waveform (of low information value), an intensity curve (in green) and an intonation contour (in white) .



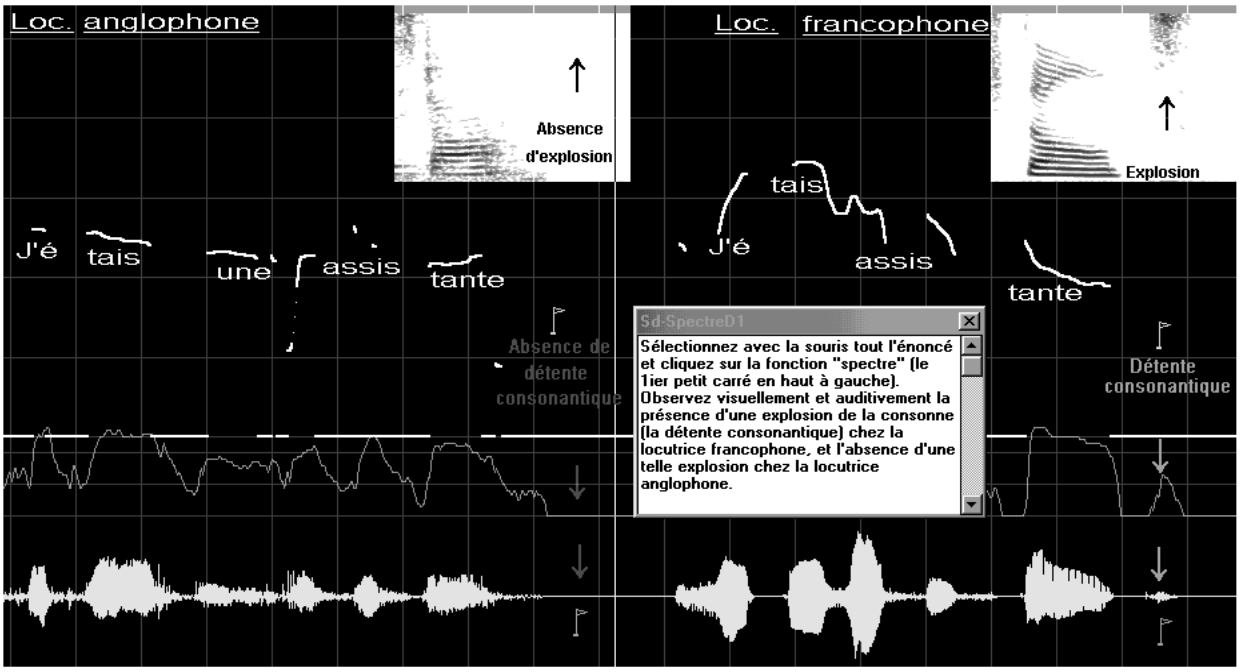
The following extracts from the Winpitch web site are quoted below:

WinPitch LTL's *Text function* (using any font available on your computer, including phonetic fonts) allows students and teachers to **place, move, and edit text** anywhere on the screen. They may use this feature: 1) to write the script of the words/sentences/texts above the melodic curve or the signal; 2) to signal a specific pronunciation or oral expression problem; 3) to comment on student's performances. Because it is so easy to move any part of the text on the screen, this writing function allows the teacher/student to **localise** the problems **precisely** and to **comment** on them.

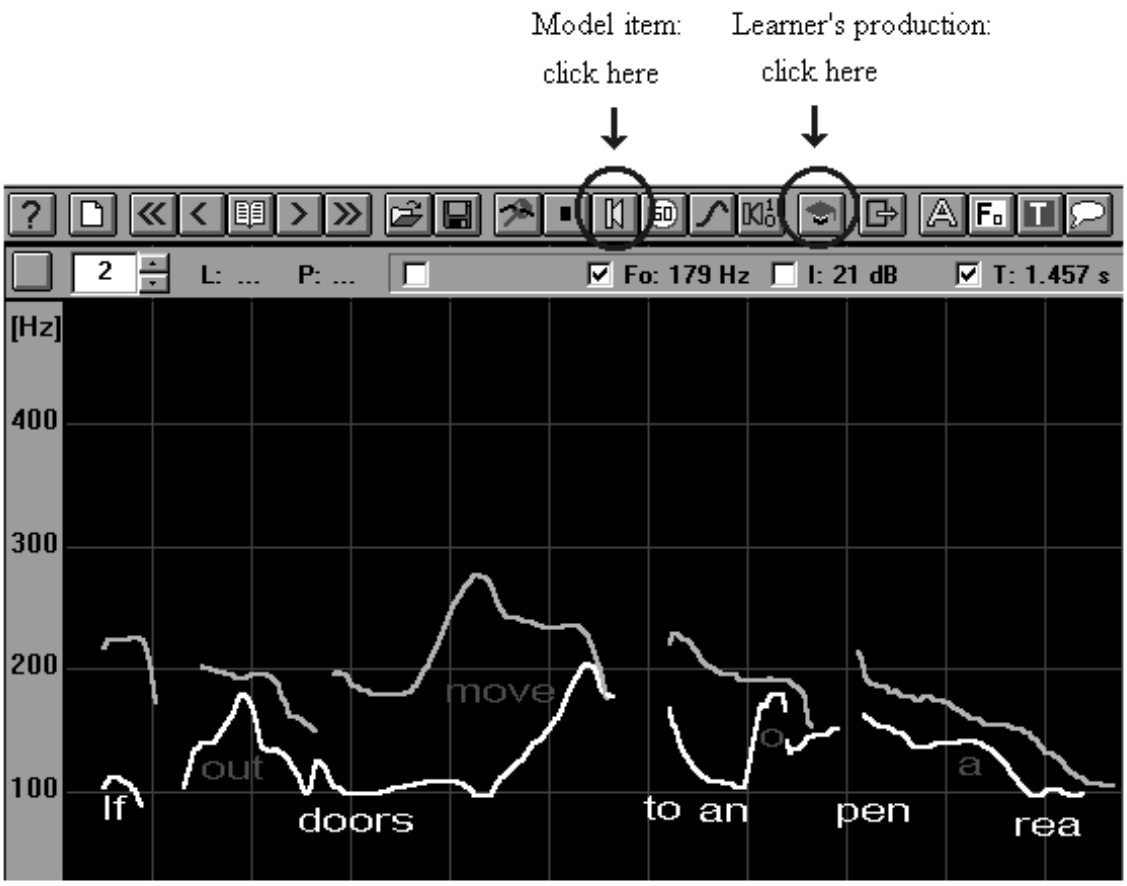
This **interactive** feature is further enhanced by **WinPitch LTL's** *Dialogue-boxes*, and also by *Bookmarks* which can be placed anywhere along the signal to allow for easy retrieval of reference points. Once again, the teacher/student can use the bookmark function and dialogue-box to localise a problem precisely, and to comment extensively on this problem without taking much space on the screen. At the touch of a key, the dialogue-box opens, permitting the student to simultaneously hear/visualise the problematic segment and read the teacher's comments.

Thus, **WinPitch LTL** enables the teacher to give **precise, extensive and individualised feedback** that links the audio and visual clues together. Just as the teacher is able to

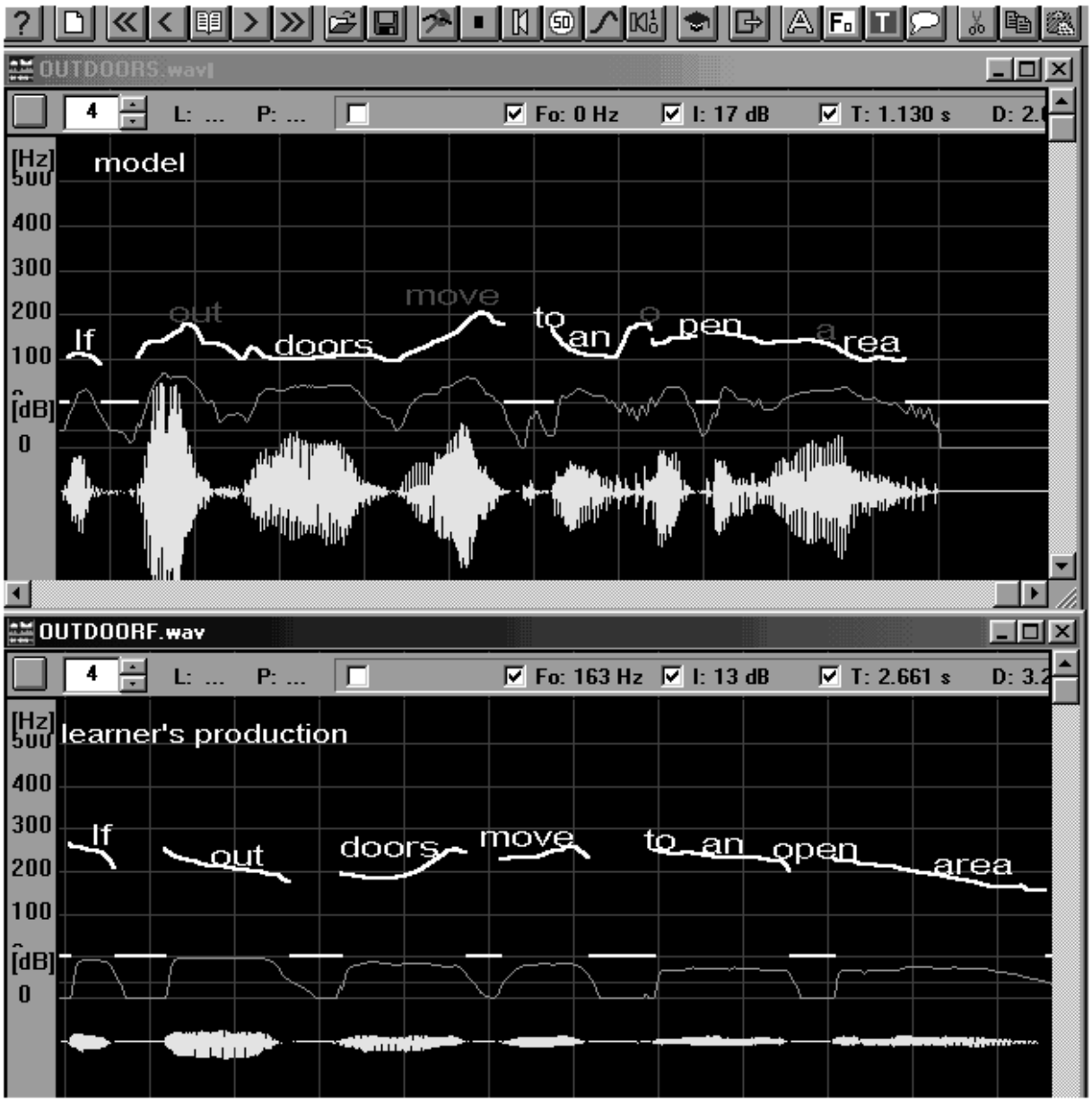
underline, highlight, comment on or correct a student's written paper, **WinPitch LTL** allows the teacher to provide the student with the same type of precise **feedback** on his/her oral production **on the computer monitor**.



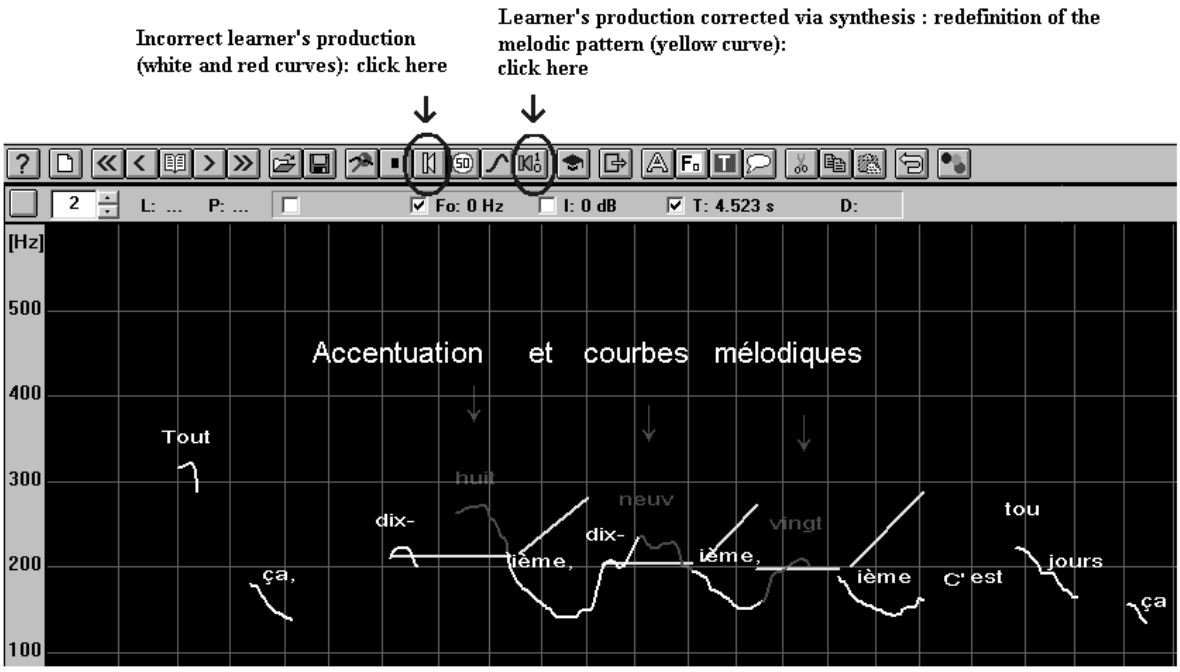
To enable the student to **compare** his/her performance to model sentences or speeches, **WinPitch LTL** provides a *Student Mode*. When this function is activated, the student, while repeating a model phrase (white curve), can visualise his/her **own melodic curve overlaying the model curve** (blue curve). **Portfolio files** allow the student to save these recordings in order to keep track of his/her progression.



To further evaluate his/her performance, the student can also use **WinPitch LTL's** *Multi Document Interface feature* to open more than one window and compare, visually and/or by ear, his/her speech production to the teacher's model.



WinPitch LTL's *Synthesis* feature allows the user to **modify the prosodic parameters** (Fo line, intensity line and duration) and to play the new defined curves by activating the synthesis mode with a simple click. With this function, the teacher is able to **redesign** a student's incorrect prosodic pattern so that the latter **can hear his/her own synthesized voice** pronouncing the correct prosodic pattern and **compare** it to his/her former speech production. The Synthesis feature functions for a whole sentence or text or block by block, enabling the student to progress segment by segment.



WinPitch LTL used in **authoring mode** allows the teacher to easily **create and organise** recorded items as well as written comments or teaching instructions into specific lessons.

Assignment 6: The final longer task

Using the information about Winpitch shown above, formulate a pedagogic strategy for teaching foreign students the basic pronunciation features of the language of your choice. The model answer given features British English.

A possible answer:

A 50 minute lesson will concentrate on 3 important issues in the learning of British English, firstly the acquisition of sounds [i:] and [i], secondly, understanding differences in word stress such as between the noun '"record" and the verb "re'cord" and finally, the mastery of essential intonation contours.

For showing the difference between "seat" and "sit", we will rely on teaching learners, formant analysis showing the difference in the look of the tense and longer vowel [i:] as against the shorter, laxer vowel [i].

The demonstration of stress is somewhat more difficult because stress perception is often a multi parameter phenomenon, stressed vowels are longer and tenser, fuller and with more intensity, both formant analysis and the intensity contour shown in blue will be used here to illustrate the difference between the noun 'record and the verb re'cord.

Intonation contours are the great strength of a tool like Winpitch, sentences of address will be taught to students to illustrate the characteristic stress and intonation contours in sequences such as "Good morning, Mrs Jones, Good afternoon, Mrs Jones, are you having a good day, Mrs Jones, etc..."

The key to engaging learners in pronunciation activities is to keep a strong focus on a single problem at a time and to keep the activities short.

All your previous assignments were for practice, this is your final assessment to be done in 3 hours

Module Final Assessment:

Short essay 1: Speech recognition (from 300 to 500 words)

Explain the way in which speech recognition has been used in CALL hitherto and what challenges still remain. Conclude by stating how you would envisage using ASR in CALL yourself, give examples of exercises and tasks which you would like to include.

Short essay 2: Speech synthesis (from 300 to 500 words)

Describe with various examples the ways in which audio and audio-visual synthesis have been used in remedial pedagogy (re-education of the deaf and CALL). What kind of future do you see for such technologies in CALL and what barriers may exist for their adoption in the foreign language classroom and digital language laboratory.

Short essay 3: Visualisation (from 300 to 500 words)

Illustrating your approach by giving examples from Winpitch, explain how you could integrate the use of visualisation in your language teaching. Do you see a place for such a tool in modern communicative pedagogy? What challenges of integration might be present? Could the tool motivate language learners? What advantages are there in improved comprehension and pronunciation for your learners?

Appendix: Further reading: The following links and references are not necessarily shown in alphabetical order

Speech Recognition

Speech Recognition – The State Of The Art
http://www.comp.mq.edu.au/courses/slp803/students/s3155370/SR.html

21st Century Eloquence Speech Recognition – Glossary of Terms
http://www.pbol.com/voice_recognition/glossary.html

Dubois, Giacomo, Guespin, Marcellesi, Marcellesi, and Mevel (eds) (1999) Dictionnaire de linguistique et des sciences du langage. Paris: Larousse.

Rodman (1999) Computer speech technology. Boston: London: Artech House.

Marcowitz (1996) Using speech recognition. Upper Saddle River, N.J.; London : Prentice Hall.

Myers (2002 in press) "Voice recognition software used for learning pronunciaiton" Proceedings of INSTILL 2000 Proceedings, InSTIL Publications.

Graham-Rowe (1999) "Read my lips." Scientific America. 14/8/99
http://www.newscientist.com/ns/19990814/newsstory8.html

Zue, Cole and Ward (1996) Ch 1.2 Speech Recognition. In Cole et ali (eds) (1996)
http://cslu.cse.ogi.edu/HLTsurvey/ch1node4.html

Cole et ali (eds) (1996) Survey of the state of the art in human language technology. CSLU http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

Aist (1999) Computer Assisted Language Learning (CALL) In Media, Design, and Applications, K. C. Cameron, ed., Swets & Zeitlinger, The Netherlands, 1999.

Ehsani and Knodt (1998) Knodt "Speech Technology in Computer-Assisted Language Learning: Strengths and Limitations of a New CALL Paradigm" In Language Language Learning and Technology Vol , No. 1, July 1998, pp. 45-60
http://llt.msu.edu/vol2num1/article3/index.html

Egan and LaRocca (2002 in press) "Speech Recognition in Language Learning: A must", InSTIL 2000 Proceedings, InSTIL Publications.

Hiller et al. (1994) “An automated system for computer-aided pronunciation learning” In CALL (1994) Vol 7, No. 1, Swets and Zietlinger, pp. 51-63

Keller and Zellner-Keller (June 2000) **New uses for speech synthesis.**
http://www.unil.ch/imm/docs/LAIP/LAIPTTS_sim.htm#uses-langteach

Speech Synthesis

The Homograph Page http://www-personal.umich.edu/~cellis/heteronym.html
TTS – State-of-the-art
http://www.comp.mq.edu.au/courses/slp803/students/s3155370/TTS.html

Goldman, Laezlinger and Wehrli (1999) La phonétisation de “plus”, “tous” et de certains nombres : une analyse phono-syntaxique. In TALN July 1999, Cargèse

ViaVoice Text-to-speech: The voice of IBM http://www-4.ibm.com/software/speech/enterprise/te_6.html

Bell Labs TTS project http://www1.bell-labs.com/project/tts

News from CTICML http://www.hull.ac.uk/cti/pubs/newsletter/dec99.htm

Jager, S. Nerbonne, J., Van Essen.J. and Van Essen, A.,(eds) "Language Teaching and Language Technology, Swets & Zeitlinger, Lisse, 1998

Keller and Zellner-Keller (June 2000) New uses for speech synthesis.
http://www.unil.ch/imm/docs/LAIP/LAIPTTS_sim.htm#uses-langteach

Dutoit (1997) An introduction to text-to-speech synthesis. Boston: Kluwer academic

Pfister and Traber (1994) Text-to-speech synthesis: An introductiion and case study. In Keller (ed.) (1994) pp. 87-107

Keller (ed.) (1994) Fundamentals of text-to-speech synthesis and speech recognition: Basic concepts, state of the art and future challenges. Paris: Larousse

Styger and Keller (1994) Formant synthesis. In Keller (ed.) (1994) pp. 109-128

Brierley and Kemble (eds) (1991) computers as a tool in language teaching. Chichester: Multilimgual matters.

Aist (1999) Speech recognition in Computer-Assisted Language Learning. In Pennington (ed) (1999)

Pennington (ed) (1999) CALL: Multimedia, Design and applications. Swets and Zeitlinger.

A-V Synthesis:

Beskow (1996) Talking heads – communication, articulation and animation. In Fonetik 96, Swedish Phonetics Conference, Nässlingen, 29-31 May, 1996

Beskow (2001) Multimodal speech synthesis. http://www.speech.kth.se/multimodal

Beskow et ali (2002 in press) Experiments with verbal and visual conversationsl signals for an automatic language tutor. InSTIL 2000 Proceedings, InSTIL Publications

Massaro and Cole (2002 in press) from 'Speech is special' to talking heads in language learning. INSTILL 2000 Proceedings, InSTIL Publications.

Lessard (1996) Introduction à la linguistique française.
http://qsilver.queensu.ca/french/cours/215/index.html

Paula Web Survey http://paula.oulu.fi/Publications/Survey/WebSurvey.pdf

The CSLU Speech Toolkit http://cslr.colorado.edu/toolkit/main.html

M. McTear. Using the CSLU Toolkit for Practicals in Spoken Dialogue Technology. In Proceedings of ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education, London, UK, Apr 1999.