Designing a Virtual Language Tutor

Preben Wik[1]
Centre for Speech Technology, Department of Speech, Music and Hearing,
KTH, Stockholm, Sweden

Abstract.

This paper gives an overview of some of the choices that have been considered in the process of designing a virtual language tutor, and the direction we have decided to take based on these choices.

Introduction

Combining computer assisted language learning (CALL), speech technology, and embodied conversational agents (ECA), is a research field still in its infancy.

In January 2004 a research project was initiated at the Centre for Speech Technology (CTT) with the objective of creating a Virtual Language Tutor (VLT), utilizing existing tools created at CTT. The target of this project is to create a piece of language learning software that contains an agent (an animated 3D-figure) that you can talk to, and that talks back to you. The tutor allows you to practice dialogues, corrects your pronunciation and pays special attention to the particular weaknesses/needs you may have. The tutor would get to know you better the more you use it, keep track of your improvements and tailor lessons based on your previous history and interaction with the tutor. As with much CALL software, the VLT can be a valuable addition to traditional classroom teaching, in that it is available when the student has time, rather than when the teacher has time, allowing for 'one-on-one' practice, and taking advantage of the computer's 'infinite patience'. In addition to this, a talking animated agent can provide the user with an interactive partner whose goal is to take the role of the human agent (Beskow et al. 2000).

Types of users

Early in the design process several types of users were considered:

- Swedish children learning English
- Adult immigrants learning Swedish
- Adult Swedes wanting to improve aspects of English (e.g. corporate English, technical English)
- Native Swedes with language disabilities wanting to improve their Swedish

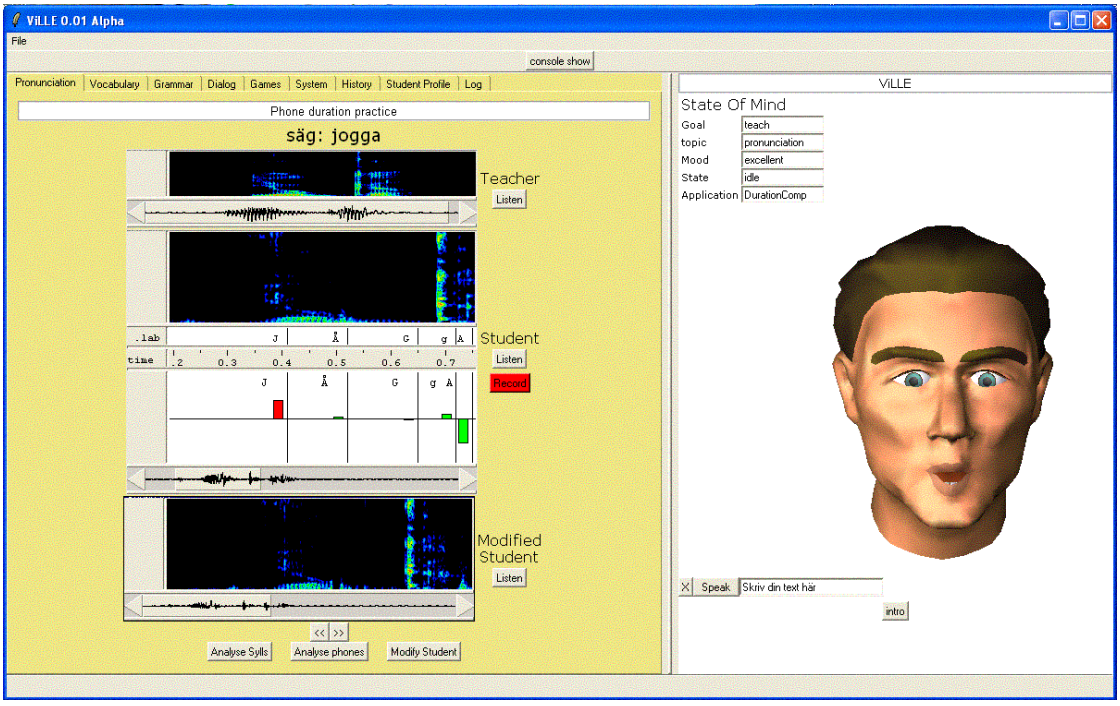

Figure 1. A screenshot of the VLT giving feedback on phone duration.

Seeing that the VLT may be useful for several groups of people with different predispositions and needs, the target from a developer's point of view is to design a system that is general enough to work for any type of user, with any linguistic background.

Separating general tools from user specific tools becomes an important issue. This way adaptation to a new user group becomes a matter of changing some user specific modules while all else can remain unchanged. Similarly, by separating linguistically universal tools from language specific ones, adaptation to a new target language will be facilitated. By keeping this distinction clear, the aim is to make a universal language tutor, with placeholders for language specific modules, and user specific applications.

A clear distinction between structure and content is also desirable, in order to allow content providers without programming skills, (e.g. language teachers) to create new and additional material for the VLT

VLT Demo Architecture

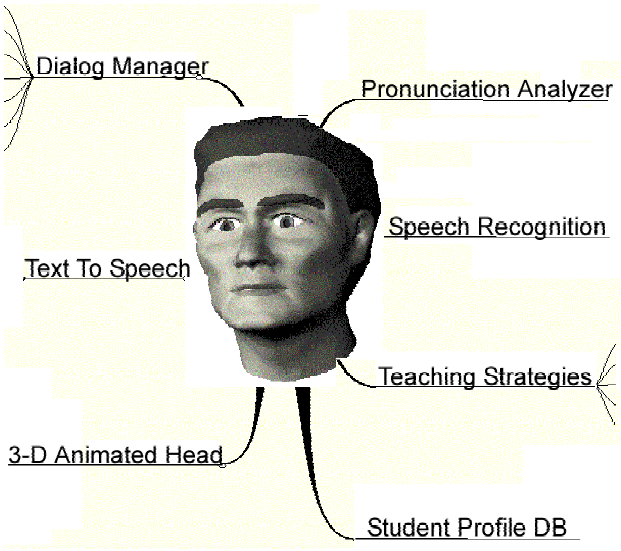

Figure 2. Schematic view of the modules that comprise the structure of the VLT.

With the universal VLT in mind, we decided to make an example application for one particular type of user, with a specific linguistic background and with a specific target language. Adult immigrants learning Swedish has been chosen as the target group. The demo will be built incrementally, adding features and wider aspects of language as the project proceeds. We decided to start our focus on detecting and giving appropriate feedback in the area of pronunciation errors.

In this scenario, the speech must both be recognized (in order to know what has been said), and later analyzed, in order to know *how* it was said. This is done by separate components. A detailed description of each element involved in the VLT is beyond the scope of this paper, but an overview of the various parts included will be given below, and can be seen in figure 2.

The VLT is built around a broker architecture. The broker receives messages from one module and passes it on to another module. This architecture allows for modules to be written in different programming languages, and the application to be distributed over several machines. The broker keeps track of which modules are loaded and what machines they are on.

Talking Head

Animated 3-D figures, i.e. talking heads have been developed at CTT (Beskow 2003). They lip-sync with synthetic or natural speech, as well as convey extra-linguistic signs such as: frowning, nodding, eyebrow movement etc.

Automatic Speech Recognition

Automatic Speech Recognition (ASR) converts speech to text. This is a difficult and error prone task. Ambiguity in the speaker's input, audio quality, the voice of the speaker etc. are factors that may cause recognition errors. An additional challenge for a VLT is that ASRs usually perform poorly on deviant speech, which is precisely what most language learners have. This poses a problem for the whole system, since the ASR is the first part of the system, where user input is evaluated. Special considerations must therefore be taken when using ASRs in a language learning environment (Witt 1999).

Dialog Manager

The Dialog Manager (DM) deals with questions such as: "What should the agent do if the user's input is X". A common way to avoid an infinite list of possible actions is to let the agent act in a restricted and limited domain.

The domain of a language teacher is possible to subdivide so that each type of exercise is viewed as a separate domain. For example, in a dialog practice 'At the restaurant', the domain will be related to dinner menus and ordering.

When practicing pronunciation the dinner menu is irrelevant.

One solution currently considered is to build many small DM agents within the agent, and let environmental variables decide which one(s) should participate in the evaluation. For example, while practicing dialogue, a mixed initiative DM is necessary. At other times, when the VLT knows what to expect from the student (e.g. in 'Say after me' types of drills), a different type of DM can be used, and more robust speech recognition can be made.

Data Base - Student Profile

A student profile database is integrated in the system. The present implementation only stores personal information such as: name, age, sex, height, and linguistic background. By logging the performance of the student as he uses the system, information that may be stored in the future include: known vocabulary, lesson history, performance history, specific pronunciation, grammatical difficulties etc.

Some other elements used by the VLT will also benefit from being stored in a database format. Tailor-made lessons can for example be produced by combining information on student-specific pronunciation difficulties, with a table of potential words to practice, connected with extra information on the phonetic characteristics of each word.

Pronunciation Analyzer

A general toolbox for pronunciation analysis is under development.

In general, pronunciation errors are likely to occur for the learner of a new language, if a distinction that carries meaning in the new language (L2), does not exist in the learner's native language (L1) (McAllister 1995, Flege 1998). The L2 learner's competence to perceive and produce difficult and new phonetic contrasts has also been shown to depend on their mother tongue (Öster 1998). A cross-reference mapping of linguistic features for each language is therefore desirable in order to make predictions about what kinds of difficulties a student is likely to have.

There are several different ways to make pronunciation errors. One often divides them into phonetic errors and prosodic/rhythmic errors.

Examples of phonetic errors are:

- Insertions - An extra vowel is sometimes inserted in certain consonant clusters
- Omissions - French people will for example often remove an initial 'h' saying 'otel' rather than 'hotel'
- Substitutions - The l/r substitution for Chinese and Japanese for example

Examples of prosodic errors are:

- Wrong vowel length - for example 'beach–bitch'
- Stress at the wrong place - for example, the word 'telephone' exists in English, Spanish, and Swedish but the stress is placed on a different syllable in each language. Eng. t<u>e</u>lephone Sp. tel<u>e</u>fono Sw. telef<u>o</u>n.

In order to automatically detect all the various types of errors, we need to first make a list of all ways to make mistakes, and then, for each item in the list, look for computational methods to analyze the signal with this error in mind.

We may end up with an incomplete list of procedures, i.e. some pronunciation errors may prove difficult to automatically detect. There could be technical reasons for failing to find a computational method, but there may also be insufficient theoretical knowledge about these errors to even know where to look for a computational solution.

Humans are able to detect minute variations in accent. A person may be considered as having a foreign accent because of minor deviations in phone quality or prosody, compared to a native speaker. Native speakers are however not a uniformly speaking group. Individual variations and local dialects may deviate more from a 'standard' than a foreign accent does, and still be considered more correct. What types of variations in speech production are perceived as normal and acceptable, is an open research area. A Nordic network (VISPP) has recently been established to address these questions.

Methods for comparing vowel length, and lexical stress, as well as good algorithms for creating a phone classifier of some kind, are currently being evaluated.

Evaluating phoneme duration was the first aspect of the pronunciation analyzer implemented in the demo. Measuring vowel length is done by using a CTT aligner tool (Sjölander 2003). It determines and time-marks phone

borders, based on a transcription of what is being said and the waveform of the utterance. The time segments are then normalized, and compared with a reference. Deviations in duration from the reference are visually displayed with rectangular bars below each phone, as seen in figure 1. If they are higher than a certain threshold they are coloured red, otherwise they are coloured green. The VLT will also give spoken feedback based on the result from the pronunciation analysis. It will either give variations of 'good', 'well done' etc. or give a remark if any phoneme duration was above the threshold.

A database of average phoneme lengths, or the rules used to determine phoneme lengths in synthesised speech, is also being evaluated as possible reference instead of pre-recorded words.

According to Bannert (1990), the single most important factor for being *understood* in Swedish, is lexical stress. We are looking at various methods to determine lexical stress in a similar way as segmental length, but looking at pitch and intensity as well as duration.

Acknowledgements

The research reported here was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

References

Bannert R. (1990) På väg mot svenskt uttal. Studentlitteratur, Lund

Bell L. (2003) Linguistic adaptations in spoken human-computer dialogues, PhD Thesis, Department of Speech, Music and Hearing, KTH, Stockholm

Beskow J. (2003) Talking Heads - Models and Applications for Multimodal Speech Synthesis. PhD Thesis. Dept of Speech, Music and Hearing, KTH, Stockholm.

Beskow J., Granström B., House D., and Lundeberg M, (2000). Experiments with verbal and visual conversational signals for an automatic language tutor. Proc of InSTIL 2000

Flege J (1998), Second-language learning: The role of subject and phonetic variables, Proc. of STiLL 199

McAllister R. (1995) Perceptual foreign accent and L2 production, Proc of the X111th International congress of Phonetic Sciences, Stockholm.

Sjölander K. (2003) An HMM-based system for automatic segmentation and alignment of speech, Proc of Fonetik 2003 Umeå University, Department of Philosophy and Linguistics PHONUM 9, 93-96

Witt S. (1999) Use of Speech Recognition in Computer-Assisted Language Learning, PhD Thesis, University of Cambridge

Öster A-M. (1998) Spoken L2 teaching with contrastive visual and auditory feedback, Proc. of ICSPL, Sydney