Design strategies for a virtual language tutor

Olov Engwall, Preben Wik, Jonas Beskow, Björn Granström

Centre for Speech Technology (CTT), KTH, Stockholm, Sweden

{olov,preben,beskow,bjorn}@speech.kth.se

Abstract

In this paper we discuss work in progress on an interactive talking agent as a virtual language tutor in CALL applications. The ambition is to create a tutor that can be engaged in many aspects of language learning from detailed pronunciation to conversational training. Some of the crucial components of such a system is described. An initial implementation of a stress/quantity training scheme will be presented.

1. Introduction

Existing computer assisted language learning (CALL) systems typically focus on the global quality of the user's phones compared to a previously defined average acoustic model. The visual feedback use waveforms and pitch curves to indicate prosody differences between the user and the model, and the "worst" word (i.e. the one with most deviant pronunciation) in the user's production is highlighted. No indication is however given as to *how* to improve the pronunciation. The student must himself identify on which phoneme the error occured, diagnose in what way his production differed from the model and understand how this could be corrected.

The requirements for a more intelligent CALL system is hence that it should: 1) identify with more precision *where* the error occurs and *what* it is. 2) keep track of its student's performance, in order to identify specific problems and adapt the exercises to address these problems. 3) give feedback that is relevant for the type of error the student made (e.g. articulatory feedback for articulatory errors). 4) give individualised feedback that indicates what features the student should practise on. 5) allow for a natural interaction with the system to practise on different aspects of language learning, from articulation training to conversations.

The CALL research at the Centre for Speech Technology (CTT) hence focuses on building a Virtual Language Tutor that addresses these issues, serving as a conversational partner, teacher and an untiring model of pronunciation, who can pick exercises from a training library depending on the user's needs. This paper discusses some of the benefits of the Virtual Language Tutor (section 2) and presents the architecture of the system in general (section 3) and the articulation training module in particular (section 4). It should be noted that the work is still at a very early stage and that the paper hence outlines our design strategies rather than presenting working components.

2. The language tutor context

Compared to current CALL systems, the use of a virtual agent has large benefits in the ability to use multimodality and gestures to give visual cues. In L2 learning, visual signals may in many contexts be more important than verbal signals and subjects listening to a foreign language often incorporate visual information to a greater extent than do subject listening to their own language [1, 2]. Conversational signals are moreover of considerable importance in the language learning context, not only to facilitate the flow of the conversation but also to facilitate the actual learning experience. We have therefore explored verbal and visual cues to signal prominence, emotion, encouragement, affirmation, confirmation and turntaking [3].

When compared to human language teachers, an automatic tutor engaged in a natural conversation still appears vastly inferior, but it does have some, at least potential, benefits over a human teacher: 1) *Pratice time*. The success of second language learning is dependent on the student having ample opportunity to work on oral proficiency. Very few human tutors have the unlimited amount of time, patience and flexibility to practise individually at any hour that a virtual tutor has. 2) *Prestige*. Many students are embarrassed to make errors in front of a human teacher, but may be less bashful about interacting with an agent. 3) *Augmented reality*. Instructions to improve pronunciation often require reference to phonetics and articulation. An agent can give feedback on articulations that a human tutor cannot easily demonstrate, by revealing articulator movements normally hidden from the outside view (cf. Fig. 1). This type of feedback may improve the learner's perception of new language sounds as well as the production by internalising the relationships between the speech sounds and the gestures.

3. System design

Considering the variety in the type of users for a virtual language tutor (e.g. both adult and child second language learners on the one hand, and speech production training of the native language for hearing-impaired children or patients with speech disabilities on the other) the aim is to design a system that is general enough to be useful for several groups of users with different linguistic background and needs. In order to achieve this, the system architecture separates the general tools from the user specific modules, linguistically

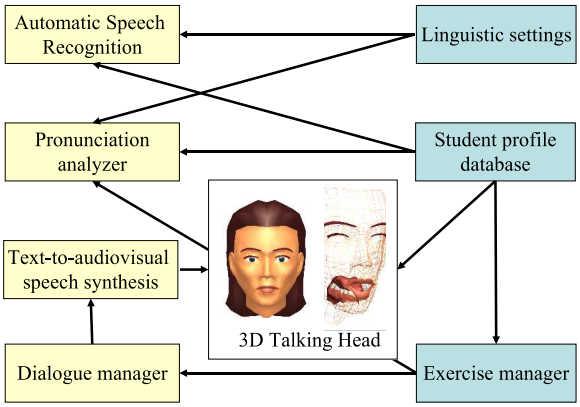

Figure 1: Overview of the modules of the Virtual Language Tutor. Left-hand side components are general system tools, while right-hand modules are adapted to the user. The arrows illustrate schematically how changes in the user-specific modules affect the general system tools.

universal tools from language specific ones and the structure from the content. This architecture makes it possible to keep large parts of the system even if a module is changed to adapt the system to a new user group, a new language or a new set of exercises.

The architecture of the Virtual Language Tutor is shown schematically in Fig. 1 and each component is described briefly in the following subsections.

3.1. Talking Heads

Several face models are available, all based on the same generic model, but personalized for the application in which it is used. Our approach is based on parameterised, deformable 3D wireframe models, controlled by rules [4]. For the purposes of animation, the parameters can be roughly divided into two (overlapping) categories: those controlling speech articulation and those used for non-articulatory cues and emotions.

The surfaces of the face can be made (semi)transparent to display the internal parts of the model. This capability of the model is especially useful in explaining non-visible articulations in the language learning situation.

The internal part includes meshes of the tongue, palate, jaw and the vocal tract walls based on the analysis of three-dimensional Magnetic Resonance Imaging (MRI) data of a reference subject [5]. Using statistical analysis, six articulatory parameters were defined to control the tongue shape.

As it is crucial for the proposed application firstly that articulations and articulatory movements are natural and secondly that the timing between the facial and tongue movements is correct, simultaneous measurements of the face (with optical motion tracking of reflective markers) and tongue (with electromagnetic articulography) movements [6] are used to train the two models in a coherent way.

3.2. Speech Recognition & Pronunciation Analyzer

The aim of the system is not only to recognize the utterances of the user, but also to detect and recognize deviations between the model pronunciation and the pronunciation of the user. This is a non-trivial extension of a standard speech recognition system; firstly because mispronounced phoneme recognition is needed, i.e. the system should be able to recognize an utterance, even if it is pronounced in a deviant way; and secondly, beacuse it should be able to locate at the phoneme level the pronunciation errors made by the speaker. These two tasks are divided into two modules in the system, the Automatic Speech Recogniser (ASR) and the Pronunciation analyzer. The role of the ASR is to transcribe the user's utterances to the system, while the pronunciation analyzer uses the output from the ASR to judge whether the pronunciation is accepted as correct or not and to spot prototypically deviant phonemes to train (i.e. finding on what part of the utterance the feedback should be focused). State-of-the-art phoneme speech recognition, with force alignment when the text is known, will be used for the ASR. Special considerations must however be taken, as the L2 learner's competence to perceive and produce difficult and new phonetic contrasts depends on the mother tongue [7]. 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. One solution to this problem is to train specific models to detect mispronounced phonemes based on the phonetic properties of both the mother tongue and the target language [8].

The Exercise manager (section 3.5) will further be used to control the focus of the training and hence which pronunciation errors – phonetic or prosodic/rhythmic – that are relevant to detect for the exercise at hand.

3.3. Dialogue Manager

CTT has a long tradition in developing multimodal dialogue systems [9] that will serve as the basis for creating different types of dialogue settings, from mixed initiative dialogues in conversation training to system prompted pronunciation drills. One solution currently considered is to build many small dialogue managers within the agent, and let environmental variables decide which one(s) to use in order to get either the most natural form of interaction, e.g. in conversation training, or the most robust speech-recognition when the expected user input is known.

3.4. Student profile database

A student profile database initially stores personal information such as name, age, sex, height and linguistic background that can be used to adapt the speech recognition and the exercises to the type of user. Subsequently, as the student uses the system, the performance will be monitored and information on known vocabulary, lesson history, specific articulatory or grammatical difficulties etc will be stored in order to provide the relevant type of training and feedback. The student's own

best production in pronunciation tasks will also be saved, to be able to use this as an alternative to the predefined reference in the feedback.

3.5. Exercise manager

The Virtual Language Tutor will be built incrementally, adding features and wider aspects of language as the project proceeds. We have currently focused on detecting and giving appropriate feedback in the area of pronunciation errors.

Evaluating phoneme duration was the first aspect of the pronunciation analyzer implemented in the demo. The CTT aligner tool [10] measures vowel length by determining and time-marking 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 signaled both by a remark from the Virtual Language Tutor and by rectangular bars below each phone in a transcription window. If the bars are higher than a certain threshold they are coloured red, otherwise green, to visualize the accepted variability. A database of average phoneme lengths or text-to-speech synthesis rules for phoneme lengths are also being evaluated as possible reference instead of pre-recorded words. We are currently looking at various methods to determine lexical stress in a similar way, but looking at pitch and intensity as well as duration.

Other exercises will be added further on and the aim is to separate the exercise manager from the technical parts of the system to allow e.g. language teachers without programming skills to add new exercises easily.

4. ARTUR - the ARTiculation TUtoR

The ARTUR project addresses another specific part of the speech training, namely the articulatory production.

A missing feature of existing systems is the possibility to establish and display a deviant articulation as a basis of constructive instruction. We therefore want to provide multimodal feedback that contrasts the user's own articulation with a correct one, using the 3D models of the face and vocal tract presented in section 3.1.

The design of this system requires a multi-disciplinary research effort and involves several tasks in addition to the modules presented in section 3. The parts that are specific to ARTUR are outlined in the following sections and in Fig. 2.

4.1. Speaker adaptation

The shape of the vocal tract varies between individuals. The articulatory model must hence adapt to each new student (scaling of the tongue, recovery of the palatal shape) to allow for correct articulatory inversion and to provide the user with visual feedback that corresponds to his or her anatomy. The adaptation can be done using medical imaging, such as MRI, to exactly scale the articulatory model to a new user, but it is of course unrealistic that every student has to be scanned with MRI before being able to use the system. We will therefore define a training procedure to establish relations between video images of the face and vocal tract dimensions. A training database of MR and facial images has been collected and computer vision techniques will be applied to extract relevant features from the video images and relate these to articulatory measures in the MR images. Based on these relations, the articulatory model can be adapted to a new student using only video images of the user's face.

4.2. Marker-less tracking of facial features

We will use computer vision analysis of a video-stream showing the face to extract information on e.g. jaw position and mouth opening. These parameters are needed both for the audio-visual speech recognition and the articulatory inversion presented in the following sections.

There are two main approaches to extracting relevant parameters. One approach is to iteratively fit a 3D model of the face to the observed face in the video images and then extract important features from the fitted 3D model [11]. Another approach is to train 2D models of face appearance conditioned on these features from a large database of face images [12]. The features can then be extracted by comparing the video images to the model without 3D reconstruction.

4.3. Audio-visual recognition of mispronounced speech

One method to increase the robustness of the speech recognition is to add visual information to the system. Neti *et al.* [13] showed that the performance of the speech recognition improved for the clean speech condition and even more so when the recognition was made with babble background noise. Correlations between jaw and lip configuration and speech acoustics [14] can further be used to link the two modalities. We will therefore incorporate visual information from the face tracking in the speech recognition to increase the robustness of the mispronunciation detection.

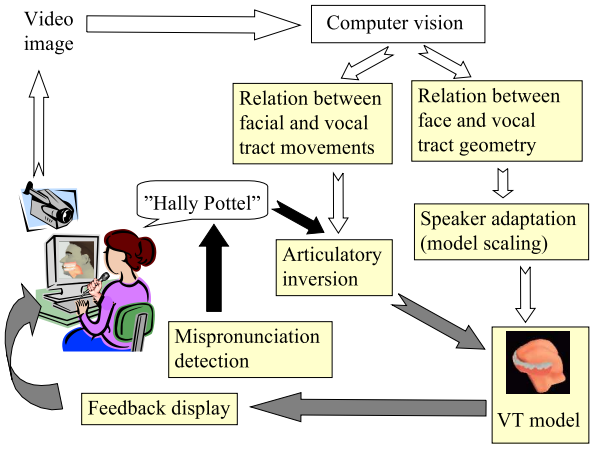

Figure 2: Overview of the components in ARTUR.

4.4. Articulatory inversion

In order to contrast the user's articulation with a correct one, the position and shape of the tongue has to be found from the speech acoustics and the visual features of the face. Neither of the two sources of information can by itself provide the information to uniquely reconstruct the vocal tract configuration. The mapping between the acoustics and the articulation is not one-to-one, several different articulatory combinations yield the same speech sound, and an acoustic-to-articulatory inversion can hence only extract candidate articulations that may have produced the acoustics. However, several studies, e.g. [15, 6] have shown that there are important correlations between 3D data of the face and the tongue position, and facial informarion will hence be exploited to guide the articulatory inversion, by ruling out candidate articulations based on the measured jaw position, lip rounding etc.

4.5. Presentation

User studies will be performed to evaluate what information is relevant to the user, and how this information should be presented to facilitate the learning. This work is performed in parallel with the technical design process and therefore requires expertise in several areas including man-machine interaction, speech therapy, pedagogy, and computer science. The development is hence made using participatory design [16] that includes all expert areas as well as the students.

In this respect, the flexibility of the talking heads used for this project is a great advantage. The articulatory feedback can be shown using a midsagittal profile with a 2D tongue contour or in 3D, showing the tongue in different reference frames (by changing the visibility or transparency of surrounding articulators), at different scales and from different viewpoints. The Wavesurfer software (http://www.speech.kth.se/wavesurfer) further provides functionality to slow down the entire utterance or parts of it, to highlight or exagerate important aspects of the articulation. We will investigate these possibilities to find what strategies are most beneficial for the users.

5. Acknowledgements

This research was carried out at the Centre for Speech Technology, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. The ARTUR project is funded by the Swedish research council.

6. References

[1] D. Burnham and S. Lau, "The integration of auditory and visual speech information with foreign speakers: The role of expectancy," in *Proc of AVSP*, 1999, pp. 80–85.

[2] B. Granström, D. House, and M. Lundeberg, "Prosodic cues in multimodal speech perception," in *Proc of ICPhS*, 1999, pp. 655–658.

[3] J. Beskow, B. Granström, D. House, and M. Lundeberg, "Experiments with verbal and visual conversational signals for an automatic language tutor," in *Proc of InSTIL*, 2000, pp. 138–142.

[4] J. Beskow, "Talking heads – models and applications for multimodal speech synthesis," Ph.D. dissertation, KTH, Stockholm, Sweden, 2003.

[5] O. Engwall, "Combining MRI, EMA & EPG in a three-dimensional tongue model," *Speech Communication*, vol. 41/2-3, pp. 303–329, 2003.

[6] J. Beskow, O. Engwall, and B. Granström, "Resynthesis of facial and intraoral motion from simultaneous measurements," in *Proc of ICPhS*, 2003.

[7] A.-M. Öster, "Spoken L2 teaching with contrastive visual and auditory feedback," in *Proc of ICSLP*, 1998.

[8] O. Deroo, C. Ris, S. Gielen, and J. Vanparys, "Automatic detection of mispronounced phonemes for language learning tools," in *Proc of ICSLP*, vol. 1, 2000, pp. 681–684.

[9] J. Gustafson, "Developing multimodal spoken dialogue systems," Ph.D. dissertation, KTH, Stockholm, Sweden, 2002.

[10] K. Sjölander, "An HMM-based system for automatic segmentation and alignment of speech," in *Proc of Fonetik, Umeå University, PHONUM 9*, 2003, pp. 93–96.

[11] J. Ahlberg, "Model-based coding - extraction, coding and evaluation of face model parameters," Ph.D. dissertation, Linköping University, Sweden, 2002.

[12] F. De la Torre and M. Black, "Robust parameterized component analysis: applications to 2D facial modeling," in *Proc of ECCV*, 2002, pp. 653–669.

[13] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition, final report from workshop 2000 audio-visual speech recognition," 2000.

[14] J. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," in *Proc of ICPhS*, 1999, pp. 199–202.

[15] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behaviour," *Speech Communication*, vol. 26, pp. 23–43, 1998.

[16] M. Muller, J. Haslwanter, and T. Dayton, *Handbook of Human-Computer Interaction*. Elsevier Science, 1997, ch. Participatory practices in the software lifecycle, pp. 255–297.