Speech Technology for Language Training and e-Inclusion

Björn Granström

CTT, Centre for Speech Technology, KTH
Lindstedtsvägen 24
SE 10044, Stockholm, Sweden
bjorn@speech.kth.se

Abstract

Efficient language learning is one of the keys to social inclusion. In this paper we present some work aiming at creating a virtual language tutor. The ambition is to create a tutor that can be engaged in many aspects of language learning from detailed pronunciation training to conversational practice. Some of the crucial components of such a system are described. An initial implementation of a stress/quantity training tutor for Swedish will be presented.

1. Introduction

Spoken language technology has already been efficiently exploited in several areas of e-inclusion. Many applications are related to the communicative needs for disabled persons. One prominent example is the early use of speech synthesis in reading machines and screen readers for visually impaired persons and as speech prosthesis for non-vocal persons. However, communication barriers to societal inclusion are not limited to persons with disabilities. In the European community, mobility is encouraged and an ambition is that every citizen should master two languages besides his/her mother tongue. The present situation in Europe is very far from this vision and a similar situation is seen in many parts of the world, limiting global communication, mobility and integration. The key to fulfilment of the ambition is naturally efficient and stimulating language learning schemes. To this end, spoken language technology has been exploited only to a limited extent. There are many possible reasons for this e.g. immature technology, lack of a sound didactic strategy and the prohibitive cost and limited spread of such technology. In this presentation we argue that the implementation of animated agents as virtual tutors in a multimodal spoken dialogue system for language training holds much promise and poses several interesting challenges for the spoken language research community.

Different agents can be given different personalities and different roles, which should increase the interest of the students. Many students may also be less bashful about interacting with an agent who corrects their pronunciation errors than they would be interacting with a human teacher. Instructions to improve pronunciation often require reference to phonetics and articulation in a way that is intuitively easy for the student to understand. An agent can for example demonstrate articulations by providing sagittal sections which reveal articulator movements normally hidden from the outside. This type of visual feedback is intended both to improve the learner's perception of new language sounds and to help the learner in producing the corresponding articulatory gestures [1]. The articulator movements of such an agent can also be synchronized with natural speech at normal and slow speech rates. Furthermore, pronunciation training in the context of a dialogue automatically includes training of both individual phonemes and sentence prosody.

2. The CTT virtual language tutor

This paper discusses some of the benefits of the Virtual Language Tutor. The research at the Centre for Speech Technology (CTT) focuses on building a tutor, using an animated talking agent, 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. In this section we also present a project aiming at an articulation training module in particular. In section 3 we present the architecture of the system in general and give an example of one application for prosody training. It should be noted that the work is still at an early stage and that the paper hence outlines challenges and work in progress that eventually may take us closer to an attractive and efficient virtual language tutor

2.1. Challenges for a virtual language tutor

Existing systems for pronunciation training typically focus on the global quality of the user's phones compared to a previously defined average acoustic model. The visual feedback uses 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 occurred, diagnose in what way his production differed from the model and understand how this could be corrected.

The requirements for a more intelligent system include:

- precisely identify the location and kind of errors.
- keep track of its student's performance, in order to identify specific problems and adapt the exercises.
- give feedback that is relevant for the type of error the student made (e.g. articulatory feedback for articulatory errors)
- give individualized feedback that indicates what features the student should practice on
- allow for a natural interaction with the system to practice all aspects of language learning, from articulation training to conversations.

While much focus needs to be on the identification and correction/display of articulation, also conversational signals with their communicative functions are of importance in the language learning context, not only to facilitate the flow of the conversation but also to facilitate the actual learning experience. It is therefore crucial that visual and verbal

signals for encouragement, affirmation, confirmation and turntaking function credibly in a multimodal system for language learning.

2.2. Providing feedback in pronunciation training

Imitation and self-correction are important factors in speech learning. While most of the pronunciation training systems that we have worked with are aimed at hard-of-hearing children, they have also been employed in adult second language (L2) teaching with very good results. Despite good hearing these students have difficulties in perceiving distinctions that are important for L2, but lacking in their L1 [2]. To use computer-assisted speech training in this situation might be particularly helpful and motivating in helping the student to significant amounts of additional training [3].

The visual feedback in the available systems is in some respect indirect since it displays articulation in terms of parameter curves, colours, dynamic maps, games etc. The parametric talking head gives possibilities to display deviant and target articulations. Some of the design issues of the general question, "How do we provide multimodal feedback that contrasts the user's own articulation with a correct one?" are described below. Should we use two displays placed side by side, showing the deviant and the model articulation? Or would it be better to display two tongues with different shapes in the same frame? Should we show just the user's tongue and highlight a place of articulation that is not reached? How should timing differences be visualized? Should we include the velum in the display in order to illustrate nasality? How should the difference between fricatives (a narrow air passage) and stops (a closed passage) be viewed?

At least five different views can be useful to illustrate differences between the student's pronunciation and the goal:

1. A frontal view where lips and other facial features can be seen clearly.
2. A side view to present the tongue movements.
3. A palatal display to show regions of contact.
4. A binary "traffic light" to show voicing.
5. Another traffic light to indicate nasality.

In this respect, the flexibility of the talking head 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, at different scales and from different viewpoints.

The WaveSurfer software we use provides functionality to slow down the entire utterance or parts of it, to highlight or exaggerate important aspects of the articulation and to change visibility or transparency of surrounding articulators in order to make the articulation as clear as possible. We will investigate these possibilities to find what strategies are most beneficial for the users. User studies show what information is relevant to the user, and how this information should be presented to facilitate 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 [4] that includes all expert areas as well as the students.

2.3. Analyzing student articulation

A prerequisite for pronunciation training with meaningful correction is a good analysis of the student behaviour and also an understanding of what benefits from correction. In ARTUR - the ARTiculation TUtoR project we address a specific part of the speech training, namely the articulatory production. [5].

The design of this system requires a multi-disciplinary research effort. The parts that are specific to ARTUR are outlined in the following sections and in Fig. 1.

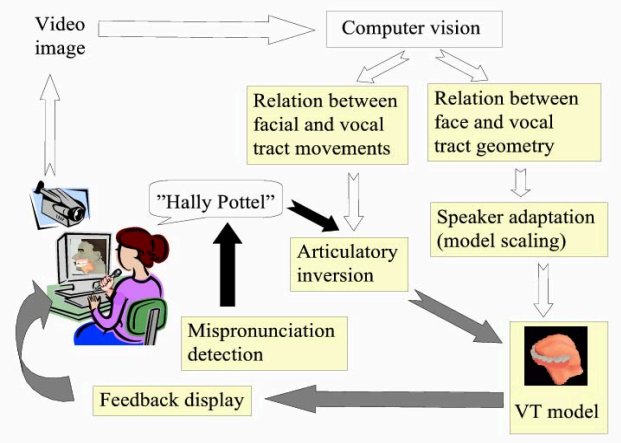

Figure 1: Overview of the components in ARTUR.

2.3.1. *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 Magnetic Resonance Imaging (MRI), to exactly scale the articulatory model to a new user, but it is of course unrealistic that every student 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.

2.3.2. *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. [6] showed that the performance of the speech recognition improved when visual information was included, especially under noisy conditions. Correlations between jaw and lip configuration and speech acoustics [7] 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.

2.3.3. *Articulatory inversion*

In order to contrast the user's articulation with a correct one, the position and shape of the tongue must 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. Thus, an acoustic-to-articulatory inversion can only extract candidate articulations that may have produced the acoustics. However, several studies, (e.g. [8,9]) have shown that there are important correlations between 3D data of the face and the tongue position. We will use facial information to guide the articulatory inversion, by ruling out candidate articulations based on the measured jaw position, lip rounding etc.

3. A first implementation

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. Massaro & Cole [10] have demonstrated the efficiency of talking heads for language training of deaf children. Bosseler & Massaro [11] have shown that using a talking head as an automatic tutor for vocabulary and language learning is advantageous for children with autism. We believe that our proposed articulation tutor will prove beneficial, aiding persons with speech production difficulties [12]. This group is very large as it includes hearing impaired children, elderly who slowly lose their articulatory precision due to a hearing impairment, patients in speech therapy and not least second language learners who have difficulties in perceiving important acoustic features of the target language. The impact of a successful implementation of a virtual language tutor is hence vast. 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 subjects listening to their own language [13,14]. Conversational signals are also of considerable importance in language learning, not only to help 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 [15].

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. Practice 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. 2). 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.

There are many potential users for a virtual language tutor, e.g. both adult and child L2 learners on the one hand, and speech production training of L1 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 all these groups of users with different linguistic backgrounds and needs. In order to achieve this, the system architecture separates the general tools from the user specific modules, linguistically universal tools from language specific ones and the structure from the content [16]. 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. 2.

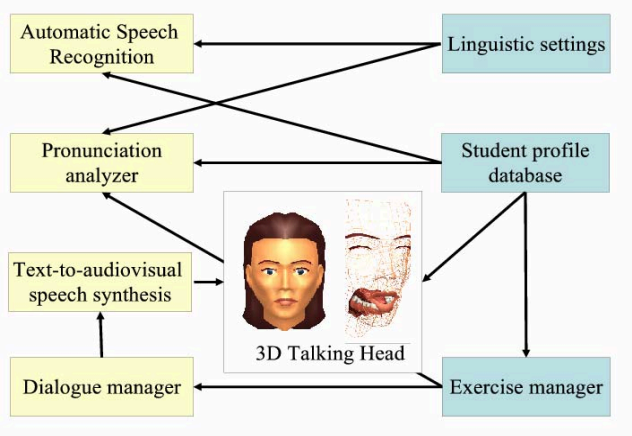

Figure 2: Modules of the Virtual Language Tutor. Left-hand side components are general system tools, while right-hand modules are adapted to the user. The arrows illustrate schematically how changes in the user-specific modules affect the general system tools.

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 recognition of mispronounced phonemes is needed, i.e. the system should be able to recognize an utterance, even if it is pronounced in a deviant way; and secondly, because 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 forced alignment when the text is known, is 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 [2]. 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 [17].
The Exercise manager will further be used to control the focus of the training and determine which pronunciation errors – phonetic or prosodic/rhythmic – that are relevant for the exercise at hand.

CTT has a long tradition in developing multimodal dialogue systems [18] 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.

3.1. Example exercise on stress and quantity

Evaluating phoneme duration is the first task of the pronunciation analyzer implemented in the prototype. The CTT aligner tool [19] 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. Feedback is supplied both by the tutor and by graphs. Deviations in duration from the reference are signalled both by a remark from the Virtual Language Tutor and by rectangular bars below each phone in a transcription window. 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. It is also possible to play a time warped version of the student utterance that conforms to the model/teacher pronunciation, thus supplying a "best pronunciation" example.

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. Limited domain conversational exercises, using dialogue systems technology are also being implemented.

4. Acknowledgements

This research was carried out at CTT, 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. The work is also dependent on several past and present EU projects including OLP, PF-STAR and CHIL This presentation relies on work of several researchers at CTT as evidenced from the references in the different sections.

5. References

[1] Badin P, Bailly G and Boë L-J, "Towards the use of a virtual talking head and of speech mapping tools for pronunciation training", In Proceedings of ESCA Workshop on Speech Technology in Language Learning (STiLL 98), 167-170, Stockholm: KTH., 1998

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

[3] Öster A-M., Hatzis A., House D., Green P., "Testing a new method for training fricatives using visual maps in the Ortho-Logo-Pedia project (OLP)", Phonum 9, Fonetik 2003, Umeå.

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

[5] Engwall O, "Speaker adaptation of a three-dimensional tongue model*" Proc ICSLP 2004*

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

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

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

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

[10] Massaro D.W. & Cole R. "From "Speech is special" to talking heads in language learning." In Integrating Speech Technology in the Language Learning and Assistive Interface, University of Abertay, Dundee 153-161, 2000

[11] Bosseler A. & Massaro D.W. (2003). "Development and Evaluation of a Computer-Animated Tutor for Vocabulary and Language Learning for Children with Autism". Journal of Autism and Developmental Disorders.

[12] Granström B (2004). Towards a virtual language tutor", Proc InSTIL/ICALL2004 NLP and Speech Technologies in Advanced Language Learning Systems, 1-8

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

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

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

[16] Engwall O; Wik P, Beskow J & Granström B, "Design strategies for a virtual language tutor", Proc of ICSLP, 2004

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

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

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