Towards a virtual language tutor

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

Abstract

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 practise. 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

The effectiveness of language teaching is often contingent upon the ability of the teacher to create and maintain the interest and enthusiasm of the student. The success of second language learning is also dependent on the student having ample opportunity to work on oral proficiency training with a tutor. The implementation of animated agents as tutors in a multimodal spoken dialogue system for language training holds much promise towards fulfilling these goals. 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 making the same errors and interacting with a human teacher. Instructions to improve pronunciation often require reference to phonetics and articulation in such a way that is intuitively easy for the student to understand. An agent can demonstrate articulations by providing sagittal sections which reveal articulator movements normally hidden from the outside. This type of visual feedback is intended to both improve the learner's perception of new language sounds and to help the learner in producing the corresponding articulatory gestures by internalising the relationships between the speech sounds and the gestures (Badin et al. 1998). The articulator movements of such an agent can also be synchronised 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.

In learning a foreign language, visual signals may in many contexts be more important than verbal signals. During the process of acquiring a language, both child L1 speakers and adult L2 speakers rely on gestures to supplement their own speech production (McNeill, 1992; Gullberg, 1998). Adult L2 speakers often make more extensive use of gestures than L1 speakers, especially when searching for words or phrases in the new language. In this context, gestures have a compensatory function in production, often substituting for an unknown word or phrase. L2 listeners may also make greater use of visual cues to aid the conversational flow than do L1 listeners. In this respect, parallels can be made between the situation of the hearing impaired listener and the L2 learner (McAllister 1998).

It has been found that the integration of segmental audio-visual information is affected by the relationship between the language of the speaker and that of the listener. Subjects listening to a foreign language often incorporate visual information to a greater extent than do subjects listening to their own language (Kuhl et al. 1994; Burnham and Lau 1999). Furthermore, in a conversation, the L2 learner must not only concentrate on segmental phonological features of the target language while remembering newly learned lexical items, but must also respond to questions at the same time. This task creates a cognitive load for the L2 listener which is in many respects much different from that for the L1 user of a spoken dialogue system. Thus, the compensatory possibilities of modality transforms and enhancements of the visual modality are well worth exploring not only concerning segmental, phoneme-level information but also for prosodic and conversational information.

2 CALL-related projects at CTT

The CALL research at the Centre for Speech Technology (CTT) focuses on building a Virtual Language Tutor, using an animated talking agent, 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. 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 Demands on a virtual tutor

Existing CALL systems fot 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 CALL system are 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 all aspects of language learning, from articulation training to conversations.

2.2 Explorative experiments on prominence

Aspects of prosody, like position and realization of phonetic quantity, stress and emphasis constitute a major problem in learning Swedish and many other languages. In human communication this is signalled not only by acoustic speech events, but rather multimodally, including facial and body gestures. In an experiment investigating the contribution of eyebrow movement to the perception of prominence in Swedish (Granström et al. 1999) words and syllables with concomitant eyebrow movement were generally perceived as more prominent than syllables without the movement. This tendency was even greater for a subgroup of L2 listeners. For the acoustically neutral test sentence the mean increase in prominence response following an eyebrow movement was 24 percent for the Swedish L1 listeners and 39 percent for the L2 group.

This type of multimodal perception of prominence can also be important for training lexical stress. In a study involving thirteen students of technical English (mostly native speakers of Swedish) the students first read a text containing words known to cause stress placement problems for Swedish speakers of English. (Hincks, 2002). The words were also selected because they often appear in technical contexts and are cognates which differ between English and Swedish primarily in the location of lexical stress. After making the recordings the students used the WaveSurfer (Beskow & Sjölander, 2000) graphical interface first to synthesize the Swedish words and then to alter the stress location from Swedish to English by manipulating duration and fundamental frequency. The students were also told to place a head nod on the correct stressed syllable. Post-test recordings made four weeks after the exercise showed a general improvement in correctly stressed syllables in the test words from about 35% (pre-test) to 70% correct (post-test).

While these experiments have addressed the issue of multimodal prominence signalling, 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.

2.3 Providing feedback in pronunciation training

Imitation and self-correction are important factors in speech learning. Most of our experience relates to the speech training of hard-of-hearing persons. Children who are born with a severe auditory deficit have a limited acoustic speech target to imitate and compare their own production with. Other senses must replace the auditory feedback that hearing children use when they learn to speak. These children seldom develop speech spontaneously, but their speech is traditionally developed through a structured training, using the visibility of speech articulation, reading, tactile sensations and, if possible, residual hearing (Ling, 1976; Dodd, 1974; Levitt & Geffner, 1987; Oller, 2000). As acoustic and visual speech are

complementary modalities, learning will be more robust and efficient than either modality alone.

In a speech training program it is important to make the student aware of the manner and place of articulation as well as of distinctive contrasts between similar speech sounds. Distinctive features like, for example, nasality/non nasality are invisible through speech reading and consequently difficult to produce correctly. A computer-assisted aid has capabilities to offer a student immediate and meaningful visual feedback of articulation and of various distinctive contrasts. By this technique it might also be easier for the therapist to instruct and explain what is wrong and what is correct in the students's production (Osberger et al, 1981; Öster, 1992).

Nowadays the use of efficient and well-designed computer-assisted speech therapy systems has been accentuated in the schools for hearing-impaired children. This kind of therapy has shown to be very efficient, especially in the instruction phase of speech training (Vicsi et al., 2000; Öster, 1996). 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 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. (Öster, 1998)

Motor learning theory in speech development indicates that accurate feedback and repeated practice are essential to establish automaticity and to transfer skills to untrained situations (Wiepert and Mercer, 2002). This is the most important element in a speech therapy program but the most difficult for a therapist to carry out. The target production must be repeated and practiced in a variety of contexts. 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 (Öster et al., 2003).

The visual feedback in the available systems are in some respect indirect since it displays articulation in terms of parameter curves, colours, dynamic maps, plays 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, or would that make the feedback too complex? 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 further 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 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 (Muller et al., 1997) that includes all expert areas as well as the students.

2.4 Analysing student articulation.

As pointed out, the necessary basis 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. (Engwall, 2004)

The design of this system requires a multidisciplinary research effort and involves several tasks. The parts that are specific to ARTUR are outlined in the following sections and in Fig. 1.

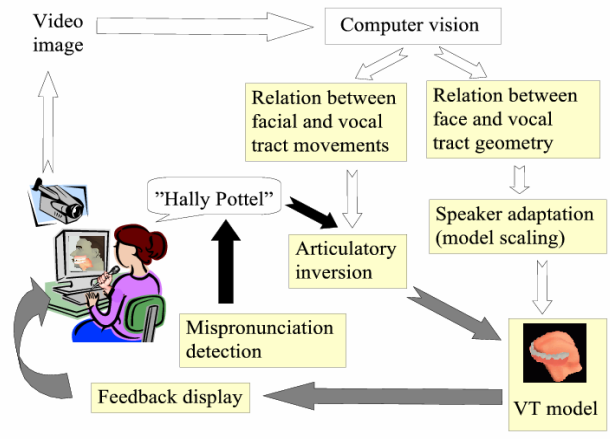

Figure 1: Overview of the components in ARTUR.

2.4.1 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 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.

2.4.2 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 (Ahlberg, 2002). Another approach is to train 2D models of face appearance conditioned on these features from a large database of face images (De la Torre & Black, 2002). The features can then be extracted by comparing the video images to the model without 3D reconstruction.

2.4.3 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. (2000) 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 (Barker & Berhommier, 1999) 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.

2.4.4 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, and an acoustic-to-articulatory inversion can hence only extract candidate articulations that may have produced the acoustics. However, several studies, (e.g. Beskow et al., 2003; Yehia et al., 1998) have shown that there are important correlations between 3D data of the face and the tongue position, and facial information will hence be exploited to guide the articulatory inversion, by ruling out candidate articulations based on the measured jaw position, lip rounding etc.

3 The CTT virtual language tutor – 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 (2000) have demonstrated the efficiency of talking heads for language training of deaf children. Bosseler & Massaro (2003) 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 as beneficial aiding persons with speech production difficulties. This group is very large as it includes hearing impaired children, elderly who slowly lose their articulatory preci-

sion due to a hearing impairment (Lane et al., 2001), 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 subject listening to their own language (Burnham & Lau, 1999; Granström et al., 1999). 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 (Beskow et al., 2000).

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.

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 universal tools from language specific ones and the structure from the content. (Wik, 2004). 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 and each component is described briefly in the following subsections.

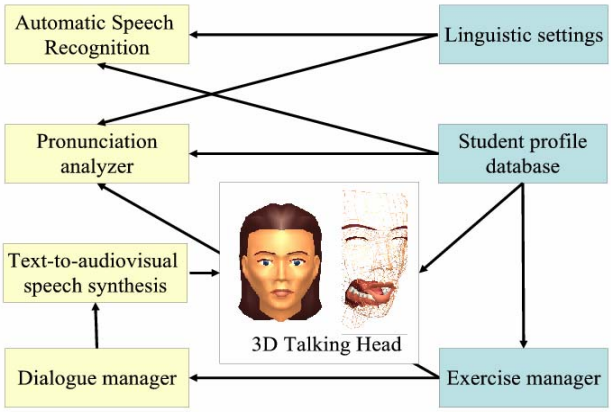

Figure 2: Modules of the Virtual Language Tutor. Left-hand side components are general system tools, while right-hand modules are adapted to the user. The arrows illustrate schematically how changes in the user-specific modules affect the general system tools.

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 (Beskow, 2003) For the purposes of animation, the parameters can be roughly divided into two (overlapping) categories: those controlling speech articulation and those used for non-verbal cues like feedback, turntaking, attitudes 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 MRI data of a reference subject (Engwall, 2003).

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 (Beskow et al., 2003) 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, 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 (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. 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 (Deroo et al., 2000).

The Exercise manager 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 (Gustafson, 2002) 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 Stress and quantity

Evaluating phoneme duration is the first task of the pronunciation analyzer implemented in the demo. The CTT aligner tool (Sjölander, 2003) 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 signaled both by a remark from the Virtual Language Tutor and by rectangular bars below each phone in a transcription window (Fig 3). 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.

The "modified student" button plays a time warped version of the student utterance that conforms to the model/teacher pronunciation, thus supplying a "best pronunciation" example.

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

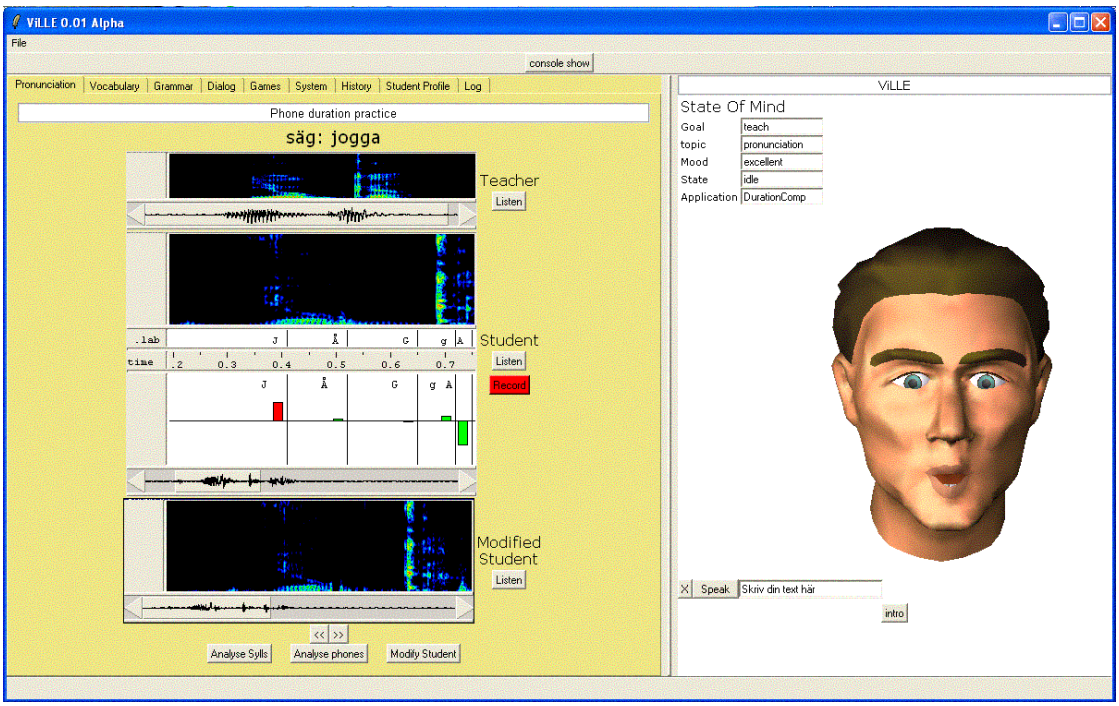

Figure 3. A screenshot of the VLT giving feedback on phone duration.

4 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. The work is also dependent on several past and present EU projects including OLP and PF-STAR. This presentation relies on work of several researchers at CTT as evidenced from the references in the different section, including Jonas Beskow, Olov Engwall, Becky Hincks, David House, Anne-Marie Öster, Kåre Sjölander and Preben Wik

References

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

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

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

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

Beskow J, 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.

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

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.

Burnham D 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.

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

Deroo O, 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.

Dodd B. (1974): The acquisition of phonological skills in normal, severely subnormal and deaf children, Doct. Diss., University of London.

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

Engwall O; Wik P, Beskow J & Granström B Design strategies for a virtual language tutor, submitted to Proc of ICSLP, 2004

Granström B, D. House, and M. Lundeberg, “Prosodic cues in multimodal speech perception,” in Proc of ICPhS, 1999, pp. 655–658.

Gullberg M “Gesture as a communication strategy in second language discourse. A study of learners of French and Swedish”, Lund: Lund University Press, 1998.

Gustafson J, “Developing multimodal spoken dialogue systems,” Ph.D. dissertation, KTH, Stockholm, Sweden, 2002.

Hincks R,” Speech synthesis for teaching lexical stress”. In Proceedings of Fonetik 2002, TMH-QPSR, 44: 153-156., 2002.

Kuhl P K, Tsuzaki M, Tohkura Y and Meltzoff A M “Human processing of auditory-visual information in speech perception: Potential for multimodal human-machine interfaces”, In Proceedings ICSLP ’94, 539-542, Yokohama, Japan, 1994.

Lane, H., Matthies, M.L., Perkell, J.S., Vick, J., and Zandipour, M. “The effects of changes in hearing status in cochlear implant users on the acoustic vowel space and coarticulation”. Journal of Speech, Language, and Hearing

Levitt H & Geffner D. (1987): Communication skills of young hearing-impaired children, ASHA Monographs, 26, 123-158.

Ling D. (1976): Speech and the hearing-impaired child: Theory and practice, The A.G. Bell Ass. for the Deaf, Inc., Washington, D.C.

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

McAllister R “Second language perception and the concept of foreign accent”, In Proceedings of ESCA Workshop on Speech Technology in Language Learning (STiLL 98). 155-158, Stockholm: KTH., 1998.

McNeill D “Hand and mind: What gestures reveal about thought”, Chicago: University of Chicago Press, 1992.

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

Neti C, 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.

Oller K. (2000): The emergence of the speech capacity, Lawrence Erlbaum Ass., Publ., Mahwah, New Jersey.

Osberger M. J. , Moeller M P & Kroese J M, (1981): Computer-assisted speech training for the hearing impaired, Journal of the Academy of Rehabilitative Audiologists, 14, 145-158.

Öster A.-M., “Spoken L2 teaching with contrastive visual and auditory feedback,” in Proc of ICSLP, 1998.

Öster A-M (1996) "Clinical applications of computer-based speech training for children with hearing-impairment". Proceedings of ICSLP-96, 4th Intl Conference on Spoken Language Processing, Philadelphia, USA, Oct 1996; 157-160.

Öster A-M. (1992) "The speech of deaf children - Phonological assessment as a basis for speech-training", Thesis work for the Licentiate Philosophy degree in Phonetics, University of Stockholm, Institute of Linguistics, April 22, 1992.

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

Sjölander K and Beskow J (2000). WaveSurfer - an Open Source Speech Tool, Proc of ICSLP2000,see also http://www.speech.kth.se/wavesurfer/

Sjölander K, “An HMM-based system for automatic segmentation and alignment of speech,” in Proc of Fonetik, UmeåUniversity, PHONUM 9, 2003, pp. 93–96.

Vicsi K., Roach P., Öster A-M., Kacic Z., Barczikay & Tantoa A., Csatári F. & Bakcsi Zs., Sfakianaki A. (2000): A multilingual teaching and training system for children with speech disorders, International Journal of Speech technology 3, 289-300, 2000.

Wiepert SL, Mercer VS. (2002) Effects of an increased number of practice trials on Peabody Developmental Gross Motor Scale scores in children of preschool age with typical development. Pediatr Phys Ther. 14:22-28.

Wik P, Designing a Virtual Language *Tutor,* Proc of Fonetik 2004

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