DiVA
http://www.diva-portal.org

Postprint

This is the accepted version of a paper published in *Computer Assisted Language Learning*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Engwall, O. (2012)
Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher
Computer Assisted Language Learning, 25(1): 37-64
https://doi.org/10.1080/09588221.2011.582845

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-52189

Journal of Computer Assisted Language Learning
Vol. 00, No. 00, Month 201x, 1–26

RESEARCH ARTICLE

Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher

Olov Engwall

Centre for Speech Technology (CTT), School of Computer Science and Communication, KTH (Royal Institute of Technology), Stockholm, Sweden

Pronunciation errors may be caused by several different deviations from the target, such as voicing, intonation, insertions or deletions of segments, or that the articulators are placed incorrectly. Computer-animated pronunciation teachers could potentially provide important assistance on correcting all these types of deviations, but they have an additional benefit for articulatory errors. By making parts of the face transparent, they can show the correct position and shape of the tongue and provide audiovisual feedback on how to change erroneous articulations. Such a scenario however requires firstly that the learner's current articulation can be estimated with precision and secondly that the learner is able to imitate the articulatory changes suggested in the audiovisual feedback. This article discusses both these aspects, with one experiment on estimating the important articulatory features from a speaker through acoustic-to-articulatory inversion and one user test with a virtual pronunciation teacher, in which the articulatory changes made by 7 learners who receive audiovisual feedback are monitored using ultrasound imaging.

Keywords: Articulation analysis, Acoustic-to-articulatory inversion, Embodied conversational agents, Articulation feedback.

Email: engwall@kth.se

ISSN: 0958-8221 print/ISSN 1744-3210 online
© 201x Taylor & Francis
DOI: 10.1080/0958822YYxxxxxxxx
http://www.informaworld.com

Introduction

Computer-animated characters are becoming more and more common in both commercial and research software targeting spoken computer-assisted language learning (CALL). There are several reasons for this, not the least that studies in human-computer interaction (Lester et al. 1997, van Mulken et al. 1998) indicate that users spend more time with the system and are more positive about its performance if they are communicating with an animated agent than with an impersonal interface. Since the amount of time spent on practice is often considered as one of the most important factors for success in second language learning (Stepp-Greany 2002), and has indeed been shown to have a high influence (Glisan et al. 1998), this alone is a very strong argument. Johnson et al. (2000) in addition argue that pedagogical animated agents have a number of advantages, such as being able to guide the learner and convey familiar conversational signals and emotions.

Animated characters are mainly used as either 1) conversational partners, who practice specific interaction situations with the learner or 2) virtual language teachers, who lead lessons on, for example, pronunciation. The paradigm for the two uses of animated characters is slightly different. In the first case, the learner enters different scenarios to interact with Embodied Conversational Agents (ECAs) in a setting familiar from communication-oriented computer games. The learner/player has a target in the interaction, such as successfully ordering a meal at a restaurant or a ticket at the railway station (Morton and Jack 2005), bargaining at a flea-market (Wik and Hjalmarsson 2009) or handling a military task in a safe and confidence-inspiring manner (Johnson and Valente 2008). The focus of this type of CALL is on spoken *communication* and success is measured in terms of task completion, with feedback often being implicit. In the simplest form of implicit feedback, the ECA does not understand, or misunderstands, if the learner's production is too deviant. Standard automatic speech recognition (ASR) of the target language can be used in this case, and non- or mis-recognitions by the ASR leads to communication breakdown between the ECA and the learner. The learner therefore gets an indication that the utterance was incorrectly produced, and is required to try again. Since no or little indication is given on what the error was; lexical, grammatical or in the pronunciation; this type of communication games with standard ASR is more suited for advanced learners, who are able to self-monitor their utterances and correct them without assistance.

If the ASR is adapted to handle non-native speech, with recognition grammars and phonetic lexicons for common learner mistakes, the feedback may be more informative. As an example, Morton and Jack (2005) use recasts, i.e., reformulating an erroneous student utterance by removing the error, in their CALL game SPELL. Recasts are commonly used in the language classrooms to signal errors without breaking the communication flow or the student's motivation by explicit corrections. However, studies on second language teaching (Carroll and Swain 1993) have shown that recasts are less effective than explicit feedback, which may be due to the fact that learners often do not get the chance to repair after recasts (Sheen 2004) or that they understand the recast as a communicative confirmation rather than a correction (Lyster 1998). There is hence a need also for explicit correction of errors in CALL, in order to ensure the effectiveness of the practice. Virtual teachers are well-suited to provide such feedback, since it would be of a type familiar from the learners' previous experiences of how human language teachers give corrective instructions.

A virtual teacher acts as a language tutor for the users, i.e., presents exercises, makes the practice more stimulating and helps the users improve by making them

aware of errors they make and ways to correct them. Since many different skills are required when learning a new language, virtual tutors have been developed to focus on different aspects, such as vocabulary (Massaro and Bosseler 2006), perception (Wik and Hjalmarsson 2009), or pronunciation on different levels, from phoneme difficulties to segmental errors (Neri et al. 2006). In this article, the focus is on tutoring to correct phonemic pronunciation errors caused by deviations in the articulation.

We have in a series of articles (Engwall et al. 2006, Engwall and Bälter 2007) argued that audiovisual feedback on how to change the articulation may be beneficial if the error was made on the articulatory level. The practice scenario is the following. The virtual tutor prompts the learner to produce an utterance in an exercise, using e.g., repeat-after-me, reading out loud of written text, quiz questions or translation of a word presented in the learner's native language or as a symbolic image. This means that the tutoring system knows the expected target and can check that the student's utterance corresponds to the target on the word level using ASR. After having established that the student uttered the intended word(s), the utterance can be automatically segmented into phonetic units through a forced alignment between the acoustic signal and written text of the words (this process is described further in the next section). The phonetic segment that is the focus of the practice is then analyzed on the articulatory level. Using hypotheses about common errors for that phonetic segment, given the student's language background, feedback on how to correct the articulation may be selected from a pool of pre-generated audiovisual instructions.

Engwall and Bälter (2007) presented a Wizard-of-Oz study of an articulation tutoring system, and it was evaluated with respect to the students' subjective impression about the interface and the training. This article follows up on that article by firstly suggesting an automatic method to determine what feedback the tutoring system should give and secondly investigating, on the articulatory level, the students' response to this type of feedback. In the first part, acoustic-to-articulatory feature inversion is proposed as a method to analyze the learner's pronunciation on the segmental level, and the performance of the method is investigated in an experiment on data-driven regression. The outcome of the feature inversion can be used to select the feedback instruction that most appropriately describes the articulatory change that the learner has to make in order to reach the target. In the second part, it is described how virtual teachers can be used to provide articulatory instructions or feedback, and a user test is presented. The goal of the user test is to investigate, by means of ultrasound imaging, if the participating learners were able to transfer the articulatory feedback instructions to their own articulation. Related previous work in these two areas of research are presented in their respective parts.

Articulation analysis

Automatic speech recognition is often used in computer-assisted pronunciation training (CAPT) to find deviations in the learner's production of phoneme targets. For articulation errors, two alternatives are available, either phoneme classification or articulatory feature detection.

The phoneme classification method is currently the most popular, since standard ASR methodology may be used. The adjustment needed is that the training is based not only on native speakers of the target language (L2), but also on second language learners, either from a specific language background (L1) or from several. This allows to set up categories of both correct and incorrect pronunciations into which the learners' input may be classified (Deroo et al. 2000, Wei et al. 2009). If an

L1 learner's pronunciation of an L2 phoneme is classified into a category defining the native pronunciation, then it is accepted. If, on the other hand, it is classified into a non-native category, an error has been detected. If the intended phoneme is known, which is often the case in teacher-led practice, articulation feedback can then be generated (Eskenazi 2009). With phoneme classification trained on speakers from only one language background, a strong hypothesis may be formed on by which L1 phoneme the target was substituted, and the feedback provided can therefore describe the articulatory difference between the L2 target and the L1 phoneme. This is however problematic, since the learner's production need not be typical for the L1 phoneme, and the actual deviation from the L2 phoneme may hence not correspond to the general differences between the two phonemes. As a consequence, the feedback may not address the change that the learner actually has to make.

Feature detection signifies that the analysis instead aims at finding aspects in the acoustic signal that can be translated into articulatory information, e.g., place and manner of articulation and lip rounding (Frankel et al. 2007, Teppermann and Narayanan 2008). The output from each detector can then be compared to the features of the target and the feedback is based on the deviating features (Strik et al. 2009). Compared to phoneme classification, the method hence has the benefit that it is possible to focus the feedback on the particular features that were not produced according to the target, rather than on general differences between phonemes. However, in pronunciation training focused on the articulation, it can be important to make a quantitative estimation of a feature, rather than only determining the class (e.g., 'front', 'central' or 'back' in the case of place of articulation). Firstly, the wording of the feedback instruction should be determined by how large the required change in articulation is. Secondly, if the student's articulation approaches the correct one, this should be acknowledged and encouraged by the virtual teacher, even if it still belongs to the wrong class.

This article therefore proposes to attempt recovering the important parts of the student's actual articulation from the speech signal, i.e. to perform a restricted acoustic-to-articulatory inversion. Pure acoustic-to-articulatory inversion signifies that the entire vocal tract configuration used to produce the speech sound is recovered. This is theoretically impossible, since the same speech sound may be produced with several different combinations of articulator positions (the many-to-one mapping) and there are too few input parameters from the speech signal to estimate the output parameters describing the vocal tract configuration (the under-determinedness). The many-to-one mapping problem can to some extent be reduced by applying constraints, such as that the speaker tries to minimize the energy or maximize the smoothness of the articulatory movements (Ouni and Laprie 2002). The problem of under-determinedness may be reduced by increasing the number of input parameters with visual information of the speaker's face, i.e., to modify the problem into an audiovisual-to-articulatory inversion (Kjellström and Engwall 2009).

As an alternative, it is in this article instead suggested that the number of output parameters may be reduced in the setting of articulation analysis for pronunciation training, by only considering the features of the articulation that are the most important for the resulting pronunciation. In the next section, methods to perform acoustic-to-articulatory inversion are first introduced. An experiment on inversion of articulatory features that may be relevant for the CAPT setting is then presented.

Acoustic-to-articulatory inversion

The problem of articulatory inversion has been attacked using analysis-by-synthesis or statistical methods. In inversion through analysis-by-synthesis, an articulatory model is used to create a codebook containing all possible articulatory feature combinations on the one hand and the resulting synthetic acoustic features on the other (Bailly and Badin 2002, Ouni and Laprie 2002). Inversion from an acoustic signal is then performed by searching the codebook for matching acoustic data and retrieving the articulatory features linked to this entry. Due to limitations in existing articulatory models, the method is often restricted to vowels, and it is hence of limited interest for pronunciation training applications.

In statistical speech inversion, a database with simultaneous recordings of acoustics and articulatory data is used to establish the relationship through machine-learning techniques. Based on the recordings, a frame-by-frame mapping is sought to recreate the articulation from the acoustic vector. As an alternative to the frame-based approach, Ananthakrishnan and Engwall (2011) instead proposed to recreate articulatory gestures (i.e., transitions between articulatory targets) from acoustic gestures. The latter method would in fact have advantages for articulation-based pronunciation analysis, since the unit 'gesture' is often more relevant when attempting to identify relative differences in the articulation. Whereas mispronunciation detection with the frame-based approach compares the articulatory positions for the target and the learner's attempt in each frame, the gesture-based approach could instead compare the direction and shape of the gesture. In this article, the traditional frame-based method is nevertheless still used.

Different regression techniques, such as linear regression, neural networks, Hidden Markov Models or Gaussian Mixture Models, can be trained with material from the database to set up a mapping between the acoustic output and the articulatory configuration that produced it.

Linear regression is straight-forward to implement and computationally cheap, and was hence used the first statistically-based inversion experiments to investigate the correlation between acoustics, the face and vocal tract data (Yehia et al. 1998, Jiang et al. 2002). The method consists in first creating a mapping matrix from training examples of combined acoustic and articulatory data. Unknown articulatory data is then estimated using the mapping matrix and the acoustics corresponding to the unknown articulation. The relation between acoustics and articulation is however non-linear, and several non-linear methods have since been applied to the inversion problem.

Artificial neural networks (ANN) have been used both separately (Kjellström and Engwall 2009) and in combination with a mixture model to create a mixture density network, MDN (Richmond 2006). The MDN has better properties for estimating continuous variables and is hence more suitable for inversion of sequences of acoustic-to-articulation data.

Context dependent Hidden Markov Models (HMM) can be used to model each phoneme with a linear estimation for every HMM state (Hiroya and Honda 2004). For each state, it is assumed that the acoustic spectrum a can be determined from the articulation vector t and the inversion consists in finding an articulation $\hat{t}$ that maximizes the a posteriori probability of observing articulation t, given acoustics a. Katsamanis et al. (2009) extended the method to audiovisual-to-articulatory inversion.

The Gaussian Mixture Model (GMM) method used by Toda et al. (2008) and Ananthakrishnan et al. (2009) approximates the acoustic and articulatory data sets as GMMs and estimates the articulatory parameters from the joint probability density of an articulatory parameter and an acoustic parameter.

The goal in the above studies was to estimate the properties of the vocal tract to the fullest extent possible, in order to establish the theoretical or statistical relationship between acoustics and articulation. The inversion method is therefore first trained on a large portion of the speech material (e.g., 90% of all frames) and then tested on the remaining part (10%), without discrimination of the phonetic content of each frame. However, for the application as articulation analysis in training with a virtual tutor, three simplifications are possible and relevant to make.

Firstly, the analysis is made only on parts that are rather stable in the acoustical signal, hence removing transitions between different articulations. This simplification is made since we are currently only able to formulate articulatory feedback on the articulator positions and not on transitions. Transitions are in fact very important for the perception and should be included in the analysis in the future, using e.g. the gesture-based inversion proposed by Ananthakrishnan and Engwall (2011). An important question that remains is then how to comprehensively describe or illustrate the differences in the correct and deviating transitions to the learner. We currently concentrate on the analysis of and feedback on articulatory positions.

Secondly, in the scope of tutor-led practice in CAPT, both the target pronunciation and common phonemic errors for that pronunciation are known. Hence, the articulation analysis for a particular target phoneme should not be trained on the entire speech material, but rather on the frames containing the target and the common mispronunciations of that target.

Thirdly, in articulation analysis it is not necessary to recover the entire vocal tract, but rather to analyze the features that are important for that phoneme. This signifies, e.g., that for the alveolar fricative [s] it is important that the position of the tongue tip is correct, while deviations in the tongue dorsum region can be accepted, since it has little effect on the pronunciation. For the velar fricative [ɧ], the situation is the opposite, with the dorsum being the important part. In the experiments described below, the inversion aims at estimating only the important part and improves the estimation by exploiting hypotheses about probable articulation errors.

Experiment on articulatory feature inversion

In the current experiment, training and testing of the articulatory feature inversion is performed between different Swedish phonemes produced by one native speaker, whereas a database of both native speakers and second language learners of Swedish would be used in a final version. The method presented here should nevertheless be relevant for articulation analysis of L2 speakers of Swedish, since it could as well be used to discriminate between a correct articulation and a deviant attempt at the same phoneme, if trained with data from both native and non-native speakers.

Data acquisition and processing

Our audiovisual database contains simultaneous video, optical motion capture, electromagnetic articulography (EMA) and acoustic recordings of one female speaker of Swedish. It was collected to provide a basis for statistical analysis of the relationship between the acoustic signal, tongue movements (captured with the EMA system Movetrack (Branderud 1985)) and facial articulation (measured in the video images or as the 3D-positions of the optical motion capture markers glued to the speaker's face and tracked by the MacReflex system from Qualisys). The experimental set-up is explained in detail in (Beskow et al. 2003, Kjellström and Engwall 2009).

The current experiment is based on two EMA coils glued onto the midsagittal plane of the tongue and four optical motion capture markers, placed at the left and right mouth corners and the upper and lower lips. One of the EMA coils was placed close to the tongue tip and the other around 30 mm further back. The frame-rate of the MacReflex optical motion capture system is 60 Hz and the EMA recordings were down-sampled to correspond to this frame rate. The articulation data t consisted of 23,409 frames with x- and y- coordinates for the two EMA coils and the x-, y- and z-coordinates of the four motion capture markers, i.e., 16 measurements in total.

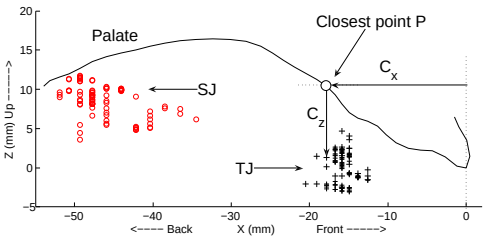
The corpus used here consisted of one recording each of 135 symmetric VCV words, where V=[a, ɪ, ʊ] and C=[p, t, k, b, d, g, f, s, ɕ, ʃ, m, n, ŋ, l, r, ɳ, ʈ, ɖ, v, j, h, jk, rk, pl, bl, kl, gl, fl, pr, br, kr, gr, kt, nt, tr, dr, fr, st, sp, sk, sl, str, spr, skr, skl] and 180 short, simple Swedish sentences, which were 3-6 words long with an "everyday content", such as "Den gamla räven var slug" (The old fox was cunning). The VCV words were designed to contain all Swedish consonants and frequently occurring or articulatorily interesting consonant clusters, all in cardinal vowel context.

The acoustic signal was originally recorded at 16 kHz, but it was divided into frames of length 24 ms with a shift of 16.67 ms to correspond to the optical motion capture frame rate of 60 Hz. Each acoustic frame was then pre-emphasized and multiplied by a Hamming window. Finally, a covariance-based LPC algorithm (Sugamura and Itakura 1986) was applied to generate 16 line spectrum pairs (LSP). The acoustic vector a consisted of 23,409 frames, each containing 16 LSP coefficients and the RMS amplitude.

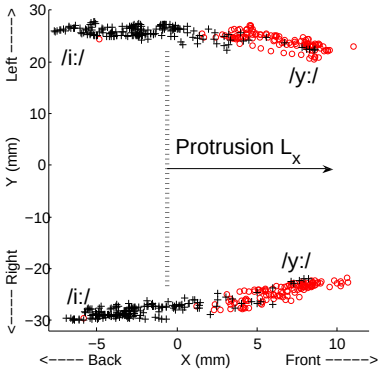
Using an HMM-based automatic aligner (Sjölander 2003), which finds the best fit between a given text string and the acoustic signal, every frame in the database was assigned a phonetic label. For each sequence of consecutive frames with identical phonetic label, the transitions from or to the adjacent articulations were removed by disregarding initial and final frames for which the second LSP coefficient differed more than one standard deviation from the mean value of that sequence. This was done since the articulation analysis is to be performed on articulator positions, rather than on transitions. It should be noted that this removal of transitions is also possible to make on-line in the CAPT situation. The second LSP coefficient was found empirically to be a good indicator for acoustically stable regions. For every phoneme p, all remaining frames labeled as belonging to that phoneme were grouped and constituted the experiment data (a_p, t_p).

As already suggested in the previous section, the goal of an articulation analysis in CAPT would be to identify the most important features of the articulation that the system can provide feedback on. When considering the articulation of the tongue, the acoustic signal is to a large extent defined by the position of the most constricted part of the vocal tract and the degree of this constriction. Lip rounding-protrusion is another feature that it would be important to provide feedback on. Finally, the vertical position of the lips is a clear articulatory difference between bilabials and labiodentals. Therefore, rather than attempting to estimate the Cartesian coordinates for the tongue coils and lip markers, the articulatory data was transformed to four relative articulatory measures (illustrated in Figure 1). They were the horizontal distance C_x between the place of constriction and the upper incisor; the vertical distance C_z between the tongue and the palate at this point; the summed protrusion L_x of the mouth corners; and the summed vertical position L_z of the lips.

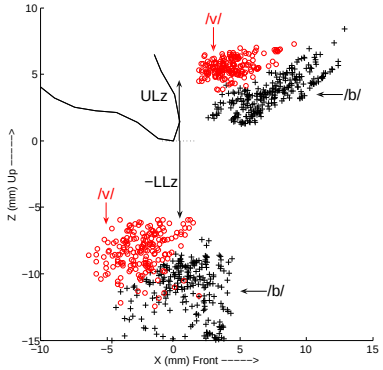
The place and degree of the linguopalatal constriction was determined as follows: Using a palate contour extracted from a Magnetic Resonance Image of the speaker,



(a) Tongue



(b) Lip protrusion



(c) Labial constriction

Figure 1. Definition of the four articulatory features (C_x, C_z, L_x, L_z) used to describe articulations and discriminate between (a) front–back and open–close (examplified by [ʃ]='SJ' vs. [ɕ]='TJ'), (b) rounded-unrounded ([i:] vs. [y:]) and (c) bilabial–labiodental phonemes ([b] vs. [v]).

the vertical distance between each of the two EMA coils and the palate was calculated for each frame. The point on the palate P with the smallest vertical distance to the tongue is defined as the place of linguopalatal constriction. For each frame, the articulation of the tongue is hence represented by C_x, the distance between P and the upper incisor, and C_z, the vertical distance between P and the tongue. This not only reduces the amount of data, but also highlights the differences between different articulations, which can be distinguished as front-back through C_x, as in Figure 1(a), or as open-close through C_z.

For the lip rounding, the horizontal protrusion L_x is calculated as the summed protrusion of the left and the right mouth corners relative to the mean protrusion over the entire corpus. That is, for a phoneme with protruded lips, such as [y:] in Figure 1(b), L_x is positive, whereas it is negative for a phoneme with spread lips, such as [i:].

To distinguish between labiodentals and bilabials, the sum of the vertical coordinates of the upper and lower lip coils (relative to the upper incisor) $Lz = UL_z - LL_z$ is used. As shown in Figure 1(c), both the upper and lower lip are higher for the labiodental [v] than the bilabial [b] and Lz is hence a good measure of the difference between these two types of labial articulations.

Restricted inversion of articulatory features

The data processing described above reduces the number of articulatory parameters that should be estimated in general. Not all these parameters are however relevant to distinguish between a target articulation and that of common mispronunciations. Depending on the phoneme, the feature set can therefore be reduced further to

only include the distinguishing features, e.g., place of articulation C_x to test the hypothesis that the velar fricative [ɧ] was produced as a front fricative, protrusion L_x to test if the lip rounding was missing in [yː], L_z to test if [b] was produced as a labiodental etc.

The next simplification concerns the training material. Since the aim is to create a detector for misarticulations of specific phonemes, a specific detector is trained for each target phoneme T and each type of probable misarticulation M. The T–M combinations are defined based on studies of common mispronunciations by second language learners of Swedish (Bannert 1994). M may consist of speech material from one as well as several phoneme classes, depending on the most probable mispronunciations for a given L1. For example, [ɑː] is commonly mispronounced as either [a] or [ɔ] by learners whose first language is Greek, and M would hence contain both these phoneme classes if the training material was from Greek speakers. If only native speakers are available in the training material, an approximation of these two mispronunciations may be created by letting M consist of native productions of [a, ɔ].

Table 1 lists the pairs of targets T and misarticulations M tested in the current experiments. These pairs include the majority of the relevant misarticulations with the tongue and the lips that occur for L2 learners of Swedish (other articulatory aspects, such as voicing, nasality and aspiration are not considered), according to Bannert (1994). A notable exception is the [l-r] distinction, since the articulatory data for the two phonemes was overlapping with the placement of EMA coils used in (Beskow et al. 2003). The T–M pairs [ɧ] *vs.* [ʊ] and [ɧ] *vs.* [k] exemplify distinctions that are relevant for the CAPT situation specifically, rather than commonly occurring pronunciation errors in non-native speech. When attempting to find the correct articulation for [ɧ], the student is likely to make the constriction too wide or too narrow and the misarticulation detector should be able to identify this.

A Gaussian Mixture Regression (GMR) method, similar to that employed by (Toda et al. 2008, Ananthakrishnan et al. 2009), was used to estimate the articulatory features from the acoustic data. The method consists in first training a set of GMMs on the joint probability density function of the articulatory features and acoustic data and then derive the mapping function from the density model. In the current Matlab implementation of the GMR (Calinon et al. 2007), the parameters of the GMMs were estimated using an Expectation-Maximization (EM) algorithm after initialization using k-means clustering. In this initialization, the number of phoneme classes N in the training material was specified. That is, if the target phoneme is contrasted against one probable misarticulation, then N=2.

For each T–M pair, a jackknife training-test procedure was employed. The jackknife procedure signifies that the articulatory-acoustic data for the frames of the T–M pair was divided into 10 equally large parts and that one part at the time was with-held from the training and instead used for testing. Through rotation, all 10 parts were used for both training and testing.

The performance of the feature inversion was evaluated by comparing the estimated articulatory features with the true ones, measured when the acoustic signal was produced. The comparison is commonly made using root mean squared error (RMSE) and/or Pearson's correlation coefficients (CC) between the estimated and true articulatory features at each time frame. In the current scope of detecting misarticulations, it is also relevant to consider the percentage of correct classifications, i.e., to what extent the articulation recreated from the acoustic signal mapped onto the articulation cluster of the correct phoneme (note that the percentage is calculated for non-overlapping parts of the original T–M clusters, i.e., if an articulation $a_M(t)$ that was in the non-overlapping region of a_M is estimated to belong to the

region shared by T and M, the estimation is considered to be erroneous).

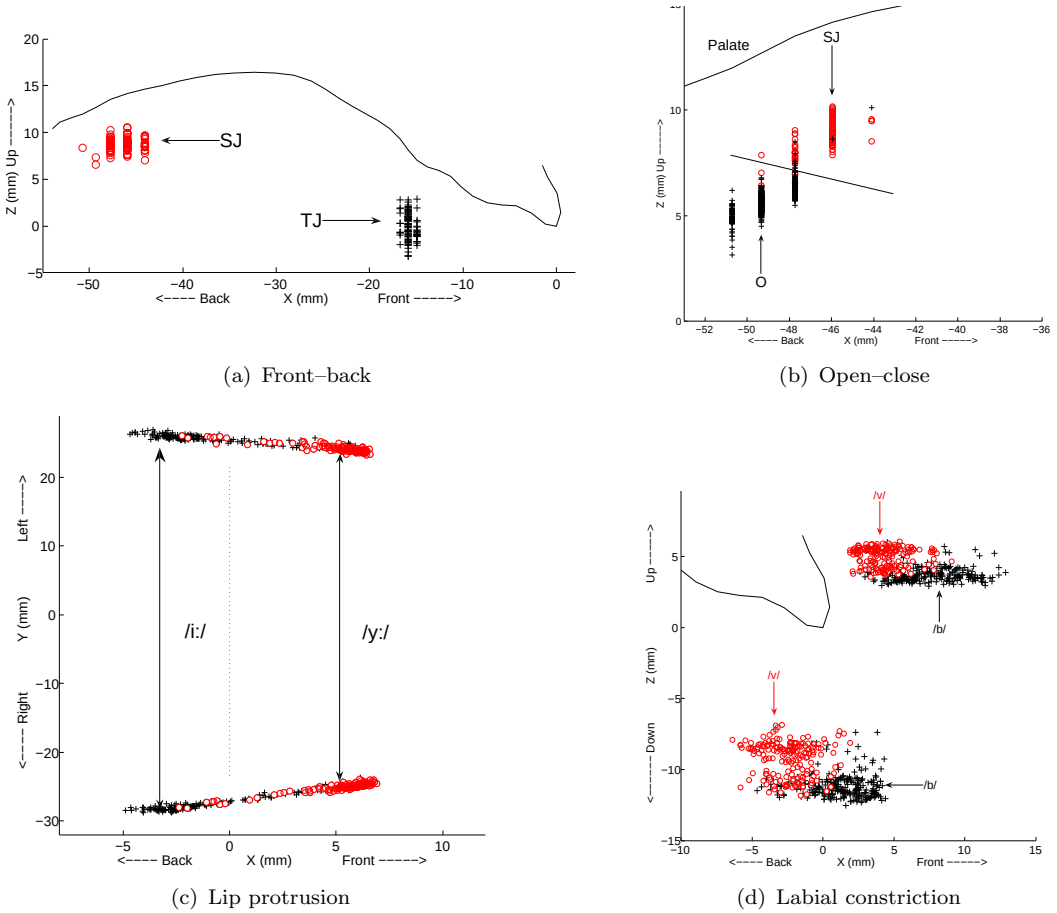
Results

Table 1 summarizes the quantitative results of the regression method and Figures 2(a-d) exemplifies the more qualitative aspects of some T–M pairs. In general, the quantitative results indicate a high level of accuracy, with correlation coefficients well above the CC=0.58 found by Kjellström and Engwall (2009) for the six parameters describing the articulation, when calculating over all frames and all phonemes in the same corpus.

The fact that the RMSE is often higher than the 2.8 mm in (Kjellström and Engwall 2009) is at least partly explained by the nature of the data. The RMSE in the current study is calculated over only one articulatory measure (place of articulation, openness, lip protrusion or lip articulation) instead of over 20 points on the tongue contour. It should also be noted that the RMSE increases with how separated the articulations of the T–M pair are, and that if the pair is well separated, the relatively high RMSE has a limited influence on the results in terms of mapping onto the correct articulation cluster. This is indicated by the high percentage of correctly classified frames, even for higher RMSE. In the scope of misarticulation detection in a CAPT setting it is more important that the qualitative aspects (i.e., which of the clusters the articulation belongs to and to what extent) are correctly estimated, than the exact articulatory feature values. This is particularly true when considering that the intended use of this feature inversion is to determine what verbal corrective feedback the virtual teacher should provide (see further the Section on Audiovisual articulatory feedback). These feedback instructions will typically ask the learner to move a specific part of the tongue "a little bit" or to change which part of the tongue that is the closest to the palate, and are hence rather vague in terms of amount of displacement (it would not be productive to ask the learner to move the back of the tongue 3.2 mm horizontally and 2.1 mm vertically).

Lingual articulation					Labial articulation				
T–M pair	example L1s	CC	RMSE (mm)	correct p %	T–M pair	example L1s	CC	RMSE (mm)	correct p %
ʃj – s	Fi, Gr, Sp	0.86	4.0	99.2	b – v	Sp	0.61	2.5	94.6
ʃj – ʃ	En	0.97	3.0	100					
ʃj – ɕ	Ge, Sp	0.98	2.6	100					
ʃj – ʊ		0.35	2.5	100					
ʃj – k		0.82	5.7	95.6					
ɕ – ʃ	En, Ar, Fr	0.67	1.8	70.1					
ŋ – j	Ar, Gr, Ja	0.74	6.4	84.8					
øː – oː	En, Ar, Ja	0.95	3.9	98.7					
ø – ʉ	En, Fi, Ja	0.88	5.0	96.5					
øː – ɛː	Gr, Ja, Sp	0.87	6.2	94.9	øː – ɛː	Ar, Gr, Sp	0.76	2.6	83.1
ø – ɛ	Gr, Ja, Sp	0.88	5.6	94.3	ø – ɛ	Ar, Gr, Sp	0.70	2.1	92.6
eː – iː	Gr, Ja	0.41	2.7	76.6	iː – yː	En, Ar, Sp	0.55	3.0	73.2
eː – ɛː	Ar, Fr, Ge	0.61	8.9	84.2					
ʉː – yː	Fi, Ge, Fr	0.53	2.1	74.8	ɔː – iː	Gr	0.75	1.6	96.5
ʏ – ʉ	Fi, Ge, Fr	0.64	2.7	85.7	ʏ – ɪ	En, Ar, Gr	0.65	1.3	96.8
uː – ʉː	En, Gr, Sp	0.94	4.9	97.9	ɑː – ɔː	Ar, Fr, Ge	0.50	4.6	100
ɑː – a	En, Ar, Fr	0.40	3.3	73.4	ɑː – a	En, Ar, Fr	0.61	1.9	86.2
	mean	0.74	4.2	89.8		mean	0.64	2.4	90.4

Table 1. *Results of the restricted inversion, in terms of Pearson's correlation coefficients (CC), root mean square error (RMSE) and the percentage of frames mapping to the articulation cluster of the correct phoneme p. For each T–M pair, examples of language backgrounds for which the distinction may be problematic according to Bannert (1994) are also listed, En= American English, Ar=Arabic, Fi=Finnish, Fr=French, Ge=German, Gr=Greek, Ja=Japanese, Sp=Spanish (the list is non-exhaustive).*



(a) Front–back (b) Open–close

(c) Lip protrusion (d) Labial constriction

Figure 2. Examples of the estimation of different articulatory features for (a) [ɧ] (o) *vs.* [ɕ] (+), (b) [ɧ] (o) *vs.* [ʊ] (+), (c) [yː] (o) *vs.* [iː] (+) and (d) [v] (o) *vs.* [b] (+).

The level of results in Table 1 should therefore be adequate for selecting the correct feedback instruction for most of the T–M pairs.

The qualitative examples in Figure 2 also show a rather adequate separation between the tested phoneme pairs and a good correspondence with the actual data in Figure 1. The main difference between the actual and the estimated data is that the inversion tends to cluster the data closer to the mean of the data set, with less variation, mainly within each phoneme, but also between the phonemes. For Figure 2(b-d) it should be noted that the slight overlap between the two classes is to a large extent not due to estimation error, but rather to overlap in the actual data (c.f., Figure 1(b-c)).

In Table 1 distinctions in either C_x or C_z have been grouped under lingual articulation features. On the other hand, differences in both tongue articulation and lip protrusion for the same T–M pair are treated separately, since they should give rise to different feedback in the CAPT setting.

Table 1 only lists phoneme pairs, but the method is also applicable to distinctions between one target and several possible mispronunciations, with similar results. For example, the inversion of the material [ɧ] *vs.* [s, ʂ, ɕ] results in CC=0.85, RMSE=3.6 mm and 99.4% correct front-back classifications. Similarly, if the material is all long or short vowels, CC=0.92 or 0.82, RMSE=6.3 or 7.9 mm and 94.9% or 95.7% correct front-back classifications, respectively.

The above experiment shows that is is possible to estimate important articulatory features to discriminate between similar phonemes that are often confounded by L2 speakers of Swedish. Two main issues should be addressed before introducing the

method in CAPT: inter-speaker variations and how the results of the articulation analysis should be presented to the learner. Since different speakers have different anatomy and different pronunciation patterns, it is necessary to find means to modify the method so that it is applicable to a group of non-native speakers, rather than one single native speaker, as in this study. This issue is briefly discussed in the Discussion. In the remaining part of this article we instead turn to how to present feedback on the learner's articulation.

Audiovisual articulatory feedback

Visual feedback directly linked to the learner's articulation has been used for quite some time in speech production training, but the measurement hardware that is required make the methods unsuitable for unassisted self-practice in second language learning.

Electropalatography (EPG), which displays the pattern of the speaker's tongue-palate contact, has long been a very useful tool in speech pathology training with in particular speech disordered children (Dent et al. 1995). The method has mainly four weaknesses for general pronunciation training. Firstly, the synthetic palate with electrodes that detect the tongue contact has to be custom-made for each individual subject, which makes it a both expensive and complicated method. Secondly, introducing a synthetic palate will affect the learner's articulation, and for second language learners, for whom the articulatory differences may be more subtle than for speech pathology patients, this is undesirable. Thirdly, since EPG only gives information about the linguopalatal contact, the variety of phonemes that can be practiced is restricted. Finally, the usefulness of EPG in speech pathology training relies on a speech pathologist who gives verbal feedback on how to move the tongue in order to approach the displayed EPG target. Should the pathologist be removed from the training situation, a very important part of the feedback is lost.

Ultrasound, which has been used with success in speech pathology training (Modha et al. 2008), has recently been proposed as a tool for pronunciation training in second language learning (Gick et al. 2008). Ultrasound has indeed benefits compared to EPG, in particular that no hardware needs to be fitted inside the speaker's mouth, thus avoiding the problem of affecting the articulation. The ultrasound probe is held or fixed under the speaker's chin, which has much less influence on the articulation and requires much less preparation. In addition, ultrasound gives images of large parts of the midsagittal tongue contour in real-time, which makes it a quite intuitive display of tongue movements. However, since ultrasound images often are very noisy and do not always show the tongue tip, they are difficult to interpret and are hence less useful for self-practice. Ultrasound in pronunciation training hence requires a speech therapist or teacher to be present to guide the learner.

Our proposition is to use a virtual teacher with an augmented reality (AR) talking head display to address the problems of these two methods. Firstly, as described in the previous part, the goal is to recover the speaker's articulation from the acoustic signal, hence avoiding interfering with the learner's speech production. Secondly, the virtual teacher can show the movements of the entire tongue with computer animations and at the same time provide explanations on the articulatory changes that are needed in order to reach the target, just as the speech pathologist is doing in speech training. This section discusses how the tongue articulation may be visualized by virtual teachers and then presents a user experiment focused on the learners' articulatory responses when they are given audiovisual feedback on

how to change their articulation. It is of interest to investigate how learners react, since both visualizations of tongue movements and feedback on how to place or move the tongue are unfamiliar to most language learners.

Articulation practice with augmented talking heads

Several previous studies on the use of augmented talking heads in pronunciation training have been performed with the models Baldi (Massaro and Light 2003, 2004, Massaro et al. 2008), MASSY (Fagel and Madany 2008) or ARTUR (Engwall et al. 2006, Engwall and Bälter 2007). All three models consist of 3D-wireframe meshes that in addition to the face also include intra-oral structures, such as the tongue, jaw and palate. In order to make the intra-oral articulations visible, parts of the face model are removed or made transparent.

Massaro and Light (2003) used four presentation conditions for Baldi: 1) a posterior view, in which the back of the head was removed, and the tongue, jaw and palate were seen from behind, together with the inside of the face texture; 2) a sagittal view displaying only the tongue and the remoter half of the palate and jaw; 3) a side view and 4) a front view, both with semi-transparent face to make the 3D structure of the tongue, palate and jaw visible. Additional information was provided by changing color of areas on the palate and teeth that were touched by the tongue and showing the airflow from the mouth for unvoiced segments. Massaro et al. (2008) used a modification of condition 2) with a cut-away side-view, displaying the midsagittal plane and the remoter part of the head, tongue, palate, velum and jaw. In order to clearly separate different articulators, they were given distinctive colors (e.g., turquoise for the tongue, green for the palate). Fagel and Madany (2008) presented the MASSY face semi-transparent in a side-view, similar to condition 3) above, but including the velum and pharynx wall. The alternative opted for in ARTUR (Engwall and Bälter 2007) is a compromise between conditions 2) and 3), as shown in Figure 3. The 3D head is presented in a side-view, but instead of making the entire face semi-transparent, the skin between the jaw and the midsagittal palate outline is made invisible. The palate itself is not displayed, to avoid that it occludes the tongue.

The advantage of the cut-away display is that the midsagittal articulatory movements and linguopalatal distances become very clear, since occluding structures are removed and the appearance is simplified. This choice is hence closely related to 2D visualizations of only the midsagittal plane, such as in SpeechTrainer (Kröger et al. 2008). The advantage of the semi-transparent display is that additional information about articulatory properties that do not appear in the midsagittal plane can be presented. Currently, no study has yet investigated either user preference or differences in learning effectiveness between the different types of displays.

The above studies have given some, but weak, evidence that subjects are able to make use of audiovisual instructions to improve their pronunciation. Massaro et al. have used Baldi to show how the tongue should be placed for different phonemes, but without giving any articulatory feedback on the learners' attempts. In each study, the learners' pronunciation was rated by listeners before and after the training.

Massaro and Light (2003) let Japanese students of English practice the pronunciation of /r/ and /l/ with either a normal front view of Baldi's face or with the four AR displays illustrating the intraoral articulation. Students improved in both training conditions, but the AR displays did not result in any additional improvement compared to the normal face view.

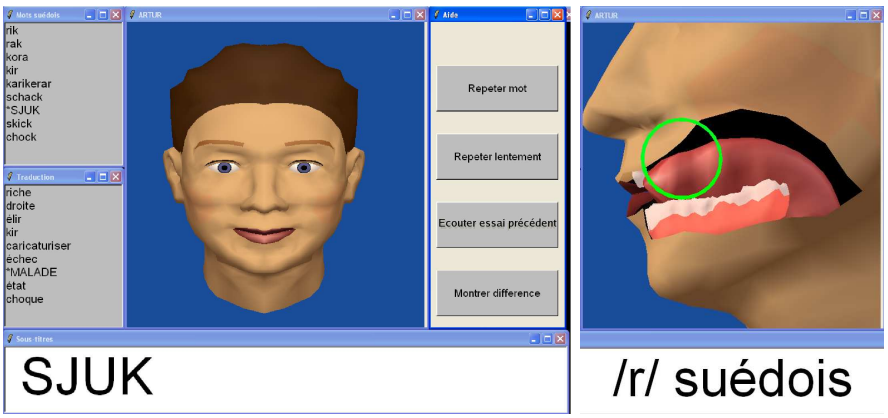
Massaro et al. (2008) gave English speakers the task to learn one pair of similar

phonemes in Arabic, with the main articulatory difference being the place of contact between the tongue dorsum and the palate. A cut-away side-view was used to illustrate the shape and position of the tongue. The improvement of the speakers' pronunciation of the two phonemes was compared in a listening test against a control group who had only been presented auditory targets, and the difference was non-significant, only 0.04 on a 7-point scale.

American hearing-impaired children who practiced consonant clusters, the fricative-affricate distinction and voicing differenced in their native language, did improve from the training with audiovisual instructions using the four AR conditions (Massaro and Light 2004). The experiment did however not include a control group and it can hence not be concluded that the children improved because they were given the additional AR information.

Fagel and Madany (2008) demonstrated the articulation of [s, z] to children with pathological lisping. The training consisted of two interactive lessons in which the experimenter used the talking head to illustrate prototypic correct articulations to the child. No feedback was given to the children on their own attempts. The pre- and post-lesson productions of [s, z] were recorded and the degree of lisping was judged by listeners. The individual differences were large, regarding both the initial status and the effect of the lessons, but the production was significantly improved after the first lesson for six of the eight children and the mean degree of lisping after two lessons was lower than pre-training for the group as a whole.

ARTUR has practiced the velar fricative [ʃj] word-initially in Swedish words with native children with pronunciation disorders (Engwall et al. 2006) and non-native speakers (Engwall and Bälter 2007). An important difference compared to the experiments described above is that articulatory feedback was given after each learner attempt to explain how the articulation should be changed. Engwall et al. (2006) and Engwall and Bälter (2007) concentrated on human-computer interaction issues and did hence not perform any formal evaluation of learner improvement, but reported that the subjective impression of the learners themselves was that the audiovisual feedback was helpful to change the articulation. In order to investigate this issue, the user test below use ultrasound imaging to monitor the learners' articulation change when presented with audiovisual feedback instructions in a short session with ARTUR.



(a) User interface

(b) Augmented reality side-view

Figure 3. *User interface for the virtual articulation teacher ARTUR. (a) Left: training words in Swedish (above) and translated into French (below), Middle: Teacher display, Bottom: instructions and subtitles, Right: user buttons ("Repeat word", "Repeat word slowly", "Listen to previous attempt", "Show difference"). The teacher display showed (a) a front face view or (b) an augmented reality side view, where parts of the cheek were invisible to show tongue positions and movements.*

Experiment on user response to articulatory feedback

The user study with seven French speakers (six male and one female) focused on the practice of the pronunciation of the Swedish alveolar trill [r] and the velar fricative [ɧ]. French was the first language for five speakers and two were bilingual with Tunisian Arabic or Persian as native language.

The aim of the user study was to analyze the articulation changes that the learners made, not to evaluate the effectiveness of the system, or of audiovisual articulatory feedback as compared to other types of feedback, such as acoustic only recasts. Instead, the expected outcome of the experiment is an indication of if the learners are able to control their tongue articulation to follow the articulatory feedback instructions. Each session lasted 5-10 minutes and it is hence very short-term changes in the articulation that are measured.

Learner interface, practice words and feedback

The interface and feedback in ARTUR was similar to the experiment presented in (Engwall and Bälter 2007), with the difference that the tutor spoke in French, with pre-recorded natural speech and time-aligned synthesized movements. Each session started with ARTUR explaining the goal of the training and introducing the interface, in particular regarding the functionality of the different user controls.

The graphical user interface consisted of four information windows and one user control frame, as shown in Figure 3(a). The central window displayed the virtual teacher, either in a normal front face view (Figure 3(a)) or an AR display showing articulatory animations (Figure 3(b)). The two progress windows to the left indicated how far into the training session the user had come, with one window for the Swedish words and one for their translation into French. The bottom window displayed sub-titles of all spoken instructions or the target word. It also prompted the user to speak by changing the background color to green. By pressing the control buttons to the right with the mouse, the learner could request an audiovisual animation of the target word at normal or slow speed, see an illustration of the difference between his own and the correct articulation, and hear an acoustic play-back of his own previous attempt. ARTUR encouraged the use of the control buttons throughout the session.

Before each student attempt, the practice word was displayed in the sub-title window and presented by ARTUR in an audiovisual animation. For the first attempt at each word, a normal view of the face (Figure 3(a)) was shown. For subsequent attempts, the augmented reality display (Figure 3(b)) was used to illustrate the tongue movements within the training word.

The training material was the Swedish words *rik* [riːk] (rich), *rak* [rɑːk] (straight), *kora* [kuːra] (select as), *kir* [kiːr] (the French aperitif), *karikerar* [karɪkeːrar] (making a caricature of), schack [ɧakː] (chess), sjuk [ɧʉːk] (sick), skick [ɧɪkː] (state) and chock [ɧɔkː]. The words were chosen so that they contained the phonemes [r] or [ɧ] in different vowel contexts (front *vs.* back, open *vs.* closed) and at the same time [k], since its palatal closure made temporal alignment between the acoustic and articulatory data possible. For [r], the position within the word was also varied.

The reason for selecting [r, ɧ] is that French instead has a rhotic [ʁ] and an alveolar fricative [ʃ], respectively. By assuming that the learners would substitute their native phonemes for these targets, corrective feedback instructions on how to change the articulation from the French to the Swedish phoneme were generated before the test. Further, the [r] articulation would be reached by primarily changing the articulation at the tongue tip, while [ɧ] involves positioning the tongue dorsum. Comparing the success in changing the articulation towards [r] or [ɧ] could hence give an indication of if the articulatory changes that the learners made were in-

fluenced by the part of the tongue that was involved. In the following description, /r/ and /ʃ/ denote the learners' attempts at producing [r] and [ɧ], respectively, without discrimination between successful and unsuccessful pronunciations.

Since the study was made before the articulation analysis described earlier in this article had been implemented, a Wizard-of-Oz-setup was used. This signifies that a human operator performs some tasks of a system without the user's knowledge, and it is a common method that allows parts or aspects of the system to be tested separately, before the entire system is ready. In this case, the human operator listened to the learner's attempt, judged the pronunciation and selected the most appropriate pre-stored feedback using short-cut keys. Feedback instructions at varying level of detail were used, depending on the progress of the training. Engwall and Bälter (2007) describe the different types of feedback used to provide encouragements, metalinguistic explanations about the spelling and elicitations to employ proprioceptic feedback or the user interaction buttons. In this article we concentrate on the corrective feedback, which was accompanied by animations illustrating the articulatory instructions in the augmented reality setting.

The corrective feedback differed between words, to maximize variation and suitability for each word, but typical corrective feedback for [r] was *"In Swedish, the /r/ sound is made with the tip of the tongue instead"*, *"Try to position the tongue tip as for /l/, but raise the edges of the tongue to the teeth instead and let the tip vibrate when air passes in the middle"* or *"Think of how you pronounce 'lake', but lower the middle of the tongue tip slightly for the first sound."* (translated from French). The accompanying animation highlighted the raised tongue tip (Figure 3(b)) or the difference compared with the French rhotic articulation.

For [ɧ], in addition to substituting the native [ʃ], it was also anticipated that learners would overcompensate and produce a uvular or pharyngeal fricative [χ,ħ] and corrective feedback was generated for the two cases. In the first, feedback was e.g., *"Pull the tongue as far back as you can, then make the constriction between the tongue and the palate"* or *"Use the same part of the tongue as for the beginning of 'cat' and lower the tongue just a little bit in the back"*. In the second, ARTUR would say e.g., *"Try to make the constriction further front, against the palate instead"* or *"The Swedish /sj/ is made against the palate, not in the throat"*. The accompanying animations either showed the articulation change suggested in the instructions (a movement backward and up towards the velum in the first case, and one forward and up in the second) or the velar place of articulation highlighted with a circle, similar to Figure 3(b).

Data acquisition & analysis

For every learner attempt, one sound file and one ultrasound image sequence were stored, in order to perform a post-analysis of articulatory changes. The ultrasound images were not shown to the learner or Wizard during the practice; all articulatory feedback to the user was through the audiovisual instructions provided by ARTUR and the Wizard judged the articulation based on the acoustic signal only.

Data of the midsagittal tongue contour was collected using a GE Healthcare Logiq5 ultrasound (US) machine and a microconvex 8C probe that produced ultrasound between 5 and 9 MHz. The possible acquisition frequency depends on the size and resolution of the image and the depth of penetration. Since subjects differed in tongue size and shape, the properties of the US acquisition set-up were adjusted individually for each subject prior to the training session to ensure that relevant parts of the tongue were captured with the highest possible image quality and frame rate. The resulting image frequency hence varied somewhat between subjects, but was around 66 Hz.

The subjects held the transducer probe under their chin with the left hand (leav-

ing the right hand free to manipulate the control buttons with the mouse). They were instructed that it was very important to hold the probe as still as possible, in order to get good measurements of their tongue movements. In quantitative speech production measurements with ultrasound it is customary to restrict the relative movements between transducer and head, since data may be lost or erroneous if the orientation of the transducer is changed. Stone and Davis (1995) therefore used a gantry to fixate both the subject's head and the transducer, while Scobbie et al. (2008) used a helmet that holds the probe. It was judged that such a set-up would interfere too much with the training situation and that it was better to let the users hold the probe. Instead, image sequences for which the position or orientation of the probe had changed too much were discarded.

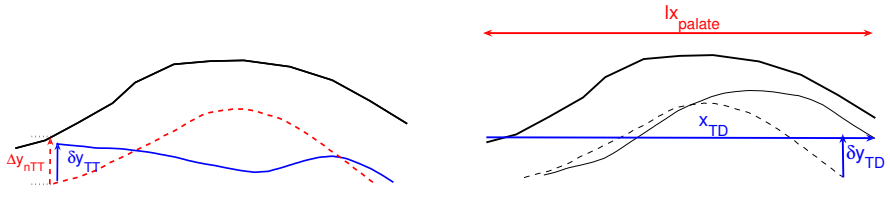
135 ultrasound image sequences were analyzed, on average 19 per subject, after discarding 27 sequences in which the tongue could not be tracked with confidence, due to probe displacement or ultrasound image artifacts. In total, 11,644 image frames were imported to the ultrasound analysis software EdgeTrak (Li et al. 2009) to semi-automatically extract the tongue contour in each frame.

The main reason for using an automatic analysis, other than a slight gain in image processing time, is that it reduces the risk of annotator bias when determining the tongue contour for different attempts. This risk is not totally avoided, since the automatic tracking has to be checked frame-by-frame for tracking errors, which are rather common in the noisy ultrasound images. Nevertheless, the active contour algorithm (Kass et al. 1987) in EdgeTrak reduces the annotator influence, since it automatically fits a curve to the air-tissue boundary at the tongue surface, by considering the image gradient and a model describing the tongue shape. This model aims at achieving intra-frame smoothness and inter-frame continuity of the contour. To reduce annotator bias in checking of the automatically extracted contours, it was made without knowledge of the acoustics that resulted from the imaged articulation, i.e., how successful the pronunciation was.

The image frames corresponding to each attempt at the target phoneme were then identified. In order to investigate the articulatory changes, the position of the tongue tip was established for each /r/ attempt and the point of maximal excursion in the tongue dorsum region was measured for each /ʃ/ attempt.

In order to examine the relative articulatory changes made by the different subjects, a normalized measure was calculated for each attempt. For /r/, the tip index was defined as the vertical distance δy_{TT} between the tongue tip position of the attempt and that of a neutral position (the average of the entire dataset), normalized by the vertical distance Δy_{nTT} from the tongue tip of the neutral tongue to the alveolar ridge (cf. Figure 4(a) for an illustration of these measures). That is, the tip index is 1 if the tip is raised to the alveolar ridge, and negative if it is lowered below the neutral tongue, as for [ʁ].

For /ʃ/, the dorsum index was defined as the horizontal position of the maximal dorsum excursion x_{TD} normalized by the horizontal length of the palate lx_{palate}.



(a) Tip index definition (b) Dorsum index definition

Figure 4. *Measures used to calculate (a) the tongue tip index and (b) the tounge dorsum index (refer to the text for the definitions). The dashed curves show the neutral tongue shape.*

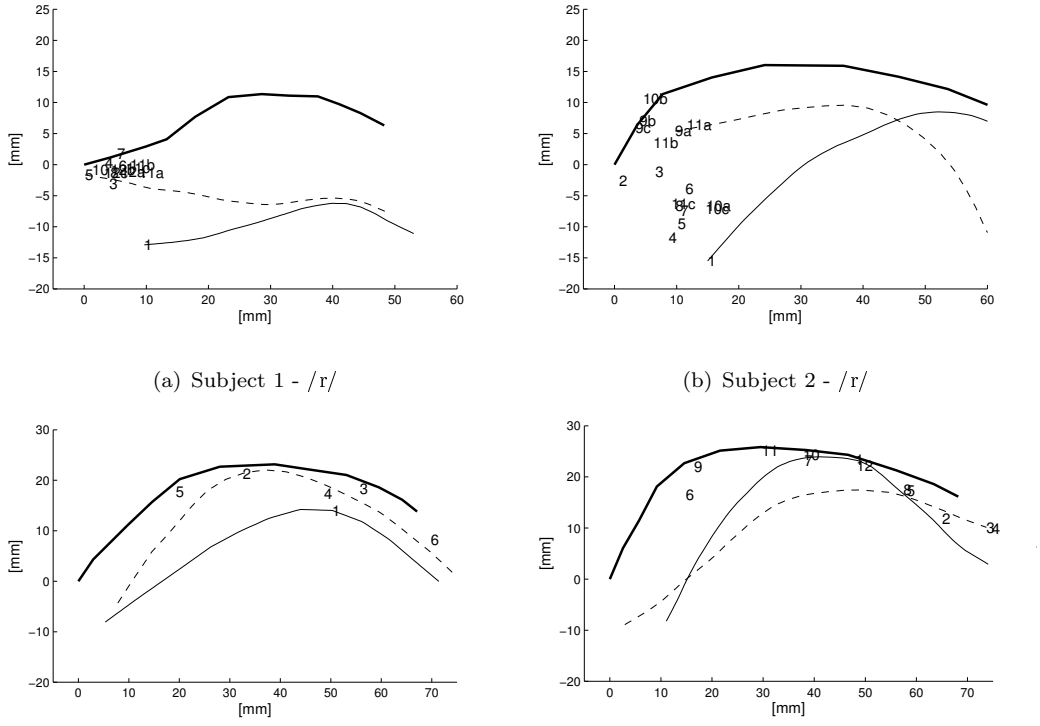
The maximal dorsum excursion was defined as the point where the tongue contour of the attempt differed the most in the vertical direction from the neutral tongue in a point-by-point comparison (δy_{TD} in Figure 4(b)). The dorsum index hence increases the further back the place of articulation is and is 1 for a velar stop (for [ʃ] it should be about 0.8, but the value depends both on the subject and the vowel context).

Since the aim of this analysis was to investigate the articulatory changes that the learners made in response to the articulatory feedback given, the acoustic signal was used only to label the articulation as resulting in on- or off-target pronunciations, not for listener ratings.

Results

Figure 5 displays examples of tongue contours and the position of the tip or dorsum in the attempts of two of the subjects for each of the two target phonemes.

The tongue tip index is shown in Figure 6. Since subjects differ in anatomy, they may achieve an alveolar articulation of [r] for different levels of the tip index, and similarly, different vowel contexts will influence the index. General trends when comparing between attempts or subjects are nevertheless visible. All subjects produced an [ʁ] articulation in the first attempt of the first word *rik*, as indicated by the negative tip index value and the off-target acoustic judgment in Figure 6. The ability to change the articulation and transfer it to new contexts differed greatly between subjects. Figures 5(a-b) & 6(a) show two different types of response to the instructions. Subject 1 rather quickly changed the articulation of /r/ to a raised tongue tip, while subject 2 reverted to producing /r/ with a lower tongue tip when



(a) Subject 1 - /r/

(b) Subject 2 - /r/

(c) Subject 3 - /ʃ/

(d) Subject 4 - /ʃ/

Figure 5. *Examples of the position of (a-b) the tongue tip in /r/ and (c-d) the dorsum in /ʃ/ in different attempts for four subjects. The solid tongue contour shows attempt 1 and the dashed line a representative example of a successful attempt. Numbers indicate the position of the tip or the dorsum in that attempt (cummulated over training words; a, b and c refer to the first, second and third /r/ in 'karikerar').*

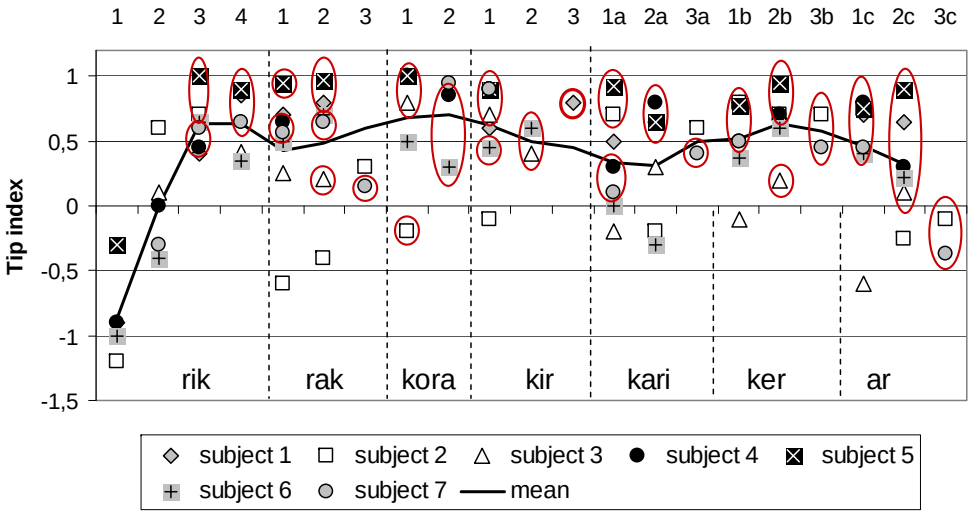


Figure 6. *The Tongue tip index of the relative position of the tip in different attempts of /r/. The mean value is calculated for attempts made by more than two subjects. Attempts labeled as correct acoustically by the annotator are circled. Attempt numbers for each training word are given at the top of the graph. Missing attempts for a subject are either due to problems extracting the tongue contour in that attempt, or that the practice had moved on to the next word.*

the training word changed. The two bilingual subjects (subjects 4-5 in Figure 6), quickly changed to an [r] articulation, which may be explained by the fact that Arabic and Persian have an alveolar [r], and these two subjects could probably replace the French [ʁ] by their own native articulation.

From Figure 6, the following observations can be made: Firstly, both as a group and for individual subjects, the tongue tip index increases with the number of attempts on the same practice word (e.g., attempts 1-4 for [riːk]), indicating that subjects approach the articulation suggested in the feedback instructions more and more. Secondly, when changing practice word, the tip index drops, because some subjects revert to the rhotic articulation in the new context (exemplified by subject 2 in Figure 5(b), as discussed above). Thirdly, the preceding vowel may influence the articulation of /r/; with higher tip index for the front close context [eːr] than for the two back open contexts [ar] in *karikerar* and for [kuːra] than for [rɑːk]. This is a natural consequence of coarticulation, but it it should be noted that (at least in this example) it is the preceding, not the following, vowel that dominates the articulation of /r/. The influence of the context is hence something that should be taken into account for both detection and feedback in CAPT.

Figure 6 also shows which attempts the annotator judged as being pronounced correctly, as judged by the acoustic signal. Mainly two types of pronunciation errors occurred; either the French [ʁ] (e.g., subject 2: attempts 1-2 in *rak*) or [l, j] following instructions that related the articulation to that of [l] (e.g., subject 6: first attempt at *ker* of *karikerar*, subject 7: attempt 2 in *rak* and 2 in *kora*). All seven subjects nevertheless changed their articulation of /r/ during the practice, using a more raised tongue tip when the audiovisual feedback instructed them to. Even if the resulting pronunciation was not always correct, it could be concluded that the learners were able to relate to the instructions concerning how to position the tongue tip.

The subjects had greater difficulties reaching the pronunciation of [ɻ], with only 9 attempts in total labeled as on-target in Figure 7. Subject 7 was discarded from the analysis, since the tongue contour could only be extracted from the ultrasound images in one attempt, due to probe movements. Similarly, only the four first

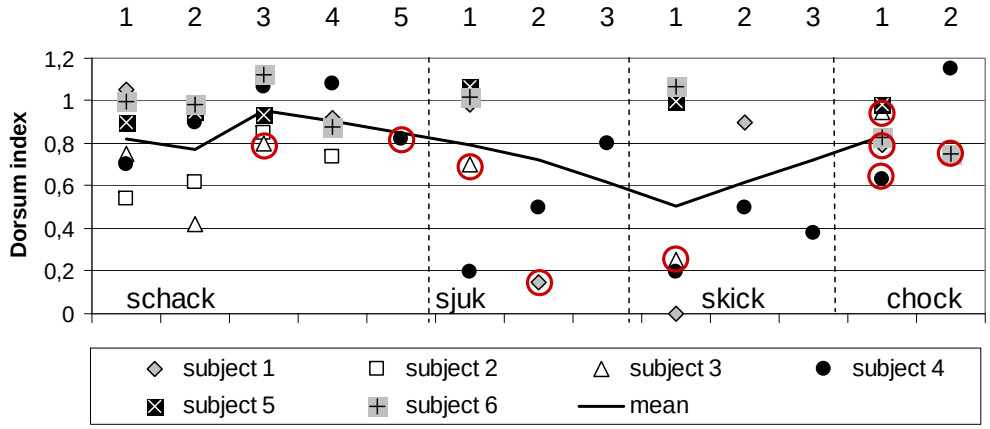


Figure 7. *(a) The Dorsum index of the relative position of the tongue dorsum in different attempts of /ʃ/. The mean value is calculated for attempts made by more than two subjects. Attempts labeled as correct acoustically by the annotator are circled. Attempt numbers for each training word are given at the top of the graph. Missing attempts for a subject are either due to problems extracting the tongue contour in that attempt, or that the practice had moved on to the next word. Subject 7 was excluded since too few attempts could be analyzed.*

attempts could be extracted for subject 2.

Due to the more complicated pattern of possible errors that occurred for [ɧ] and the smaller articulatory difference between the target and the mispronunciations, Figure 7 does not show the clear trends that could be observed for [r]. The errors that occurred were that the constriction was too far back [χ], too constricted [k] or too open [ʊ]. The dorsum index is higher for both [χ, k] and lower for [ʊ] than for [ɧ] in the same context, but because of coarticulation the value can differ substantially for a correct pronunciation (as illustrated by the low dorsum index for subject 1: attempt 2 at *sjuk* and subject 3: attempt 1 at *skick*).

There are at least four possible explanations for the subjects' difficulties to master [ɧ]: 1) Whereas [r] exists in many languages that the learners had probably encountered before the practice, [ɧ] was most probably an unknown articulation for them. 2) It is plausible that the subjects found it more difficult to consciously control the tongue dorsum than the tip, since the alveolar place of articulation for [r] is common to many familiar phonemes and the awareness of the tip movements is generally higher. The above two issues should be handled by an awareness of the difficulty of the task, providing the learners with additional varied exercises and feedback. On the other hand, there are more direct consequences for the feedback for the following two. 3) The difference in spelling clearly influenced the speakers, who frequently produced a [k] for *schack* and *skick*. This problem may be alleviated by metalinguistic feedback. In the current test it was explained that the sound was the same, even if the spelling had changed. Other alternatives could have been to present the acoustic target without the written text initially, to avoid the influence of the spelling, or to present feedback using a simplified spelling of the training word (e.g., *"Try to pronounce the word as if it was spelled /sjac/."* for *schack*). 4) The assumption that the subjects would initially use the palato-alveolar fricative [ʃ] did not hold. Almost all subjects instead replaced the fricative by the velar stop [k] or a closed back vowel [u]. As a consequence, the pre-generated feedback that instructed the learner to pull the tongue as far back as possible, or to start from a [k] articulation and release to create the fricative was not optimal. In order to improve the feedback, it should cover all possible articulation changes that may be relevant, not only the ones that are most probable from a linguistic perspective (e.g., in the case of [ɧ], instructions to change the articulation from [k] and from

[u] are also required).

In addition to the above problems for the learners, it was also more difficult for the Wizard to discriminate between the velar and the uvular fricative in real-time and some false accepts occurred, depriving the subject of a chance to correct (subject 5: attempts at *sjuk, skick, chock*). Despite the problems the subjects had in reaching the pronunciation of [ɧ], their articulatory changes are still relevant, since this study concerns how learners respond to articulatory instructions. Therefore, it is a positive indication that the subjects are aware of how to change the articulation not only when they are able to achieve the correct pronunciation, but also when they follow the articulatory instructions, even if the acoustic result is incorrect in the short-term. Examples of changes that are successful in both articulation and acoustics are when subjects produce [ɧ], either following a uvular fricative and feedback instructions to make the constriction further front (subjects 1 & 4: attempt 2 at *sjuk*, subject 6: attempt 2 at *chock*), or following a back vowel and instructions to raise the back of the tongue slightly (subject 1: attempt 2 at *skick*, subject 3, attempt 3 at *schack*). An example of a change that could be said to have been successful in the articulation only is when subjects are producing a uvular fricative after having been instructed to pull the tongue as far back as possible (subject 4: attempts 2-4 at *schack*).

Discussion

The articulatory feature inversion method presented gave some quite promising results of how different phonemes, or, in the case of CAPT, correct and incorrect pronunciations, can be discriminated in the articulatory space, using the acoustic signal as input. The experiment was performed off-line and was trained and tested on the same native speaker (which is the current state-of-the-art for data-driven speech inversion), whereas CAPT software would have to perform the analysis in real-time, for speakers that it has not been trained on, and who have a deviant pronunciation. The signal processing (identifying the frames that constitute an attempt at the target phoneme, removal of articulatory transitions) and the estimation of features from the sound signal could easily be performed in real-time, once the inversion method has been trained on an adequate set of L2 speakers. What remains to be solved is to automatically and efficiently extend the method to more than one speaker, by adaptation of features to each new speaker and/or making the method speaker independent. In general, articulatory models may be adapted in size through vocal tract length normalization based on acoustic features, such as fundamental frequency F0. Speaker independence is achieved by training the method on a large number of speakers. In this case, a combination of adaptation and speaker independence is probably required. Either, all speakers used for training are first adapted to a common articulatory space before training and the test speaker (the CAPT student) is then adapted to this common articulatory space. Or, each speaker used for training generates one individual articulatory space and the test speaker is then evaluated against the training speakers with the most similar vocal tract anatomy, as estimated by the acoustic features. While the latter method has the benefit of avoiding adaptation of the articulatory spaces, it requires more speakers for training in order to cover both the anatomical and the L1 diversity. Since the presented method is based on relative features to discriminate between different articulations, rather than an absolute inversion of Cartesian coordinates, it is however less vulnerable to a change of speaker.

Another alternative would to base the articulation analysis on gesture inversion (Ananthakrishnan and Engwall 2011), since this would further reduce the problem

of variability in speaker anatomy, because gestures can be analyzed relative to the speaker's own articulatory space. As an example, consider the difference between the Swedish and French articulation of /r/ described above and the sequence *"ara"* (e.g., in Swedish *"paraply"* or French *"parapluie"*, both meaning "umbrella"). The feature values describing place and the degree of constriction would differ substantially between speakers that all produce the same phoneme ([r] or [ʁ]) and it can therefore be difficult to define the limit for a correct articulation. There is, on the other hand, a qualitative difference between the gestures for [aʁa] (the tongue dorsum moving back and then forward) and for [ara] (the tongue tip moving upwards and then down). To discriminate between the two articulations one can therefore analyze the direction of the gesture, and it would be largely common for all speakers producing the same phoneme.

Regardless of the method used, more variability will be introduced when the method is applied to more than one speaker and discrimination between different articulations will be made with less confidence. This will affect the feedback that can be provided, so that the articulatory differences between the target and the deviant production will have to be larger. However, the type of feedback that is currently provided by the virtual teacher in itself assumes larger differences, since both the spoken instructions and the animations would be difficult to follow if the articulatory change that is needed is too small. The application hence primarily addresses practice where larger deviations can be expected and resorts to an articulatory description of the target if it is not possible to determine the change that the learner needs to make.

The work on the articulatory feature inversion is foreseen to continue with recordings of an acoustic-articulatory database of both L1 and L2 speakers of Swedish. The database will be used to analyze mispronunciations on the articulatory level and to train and test the inversion method. As the database will contain several different speakers, instead of just one as here, it will be possible to study the inter-speaker differences that are due to the anatomy and attempt to adjust for these differences, as outlined above. As the database will contain non-native speakers, it will be possible to find empirically, rather than theoretically as here, the mispronunciations that occur for a given phoneme and hence adjust the training material for the articulation analysis of that phoneme. The database will further allow us to evaluate the method on a material that is more similar to the task in CAPT. The current evaluation on one native speaker is restricted, but it does show the possibility of identifying mispronunciations on the articulatory level, which is a prerequisite for providing relevant feedback on the articulation.

The user study on the uptake of articulatory feedback gave some evidence that the learners were able to change their articulation according to the instructions on how to re-position the tongue. Massaro and Light (2003) and Massaro et al. (2008) have previously found that learners in their studies did not improve significantly more than the control groups that were only presented an acoustic target or a normal face view. Since the experimental set-up and the scope of the current study was different (no control group; evaluation of learner change performed on the articulatory level, rather than on the pronunciation; no pre-test was performed and learners may in some cases have mastered the target already; the evaluation was performed during the training, not in a post-test) it is not possible to claim that learners in the current study were more successful at learning the pronunciation. There is however a possibility that there is a relevant difference in effectiveness between providing general articulatory instructions and feedback related to the learner's attempts, since self-correction of the articulation is easier than imitation. This possibility should be investigated further.

The problems that the subjects had when attempting to produce [ɧ] further highlights a number of issues that are relevant for the practice of other phonemes as well, namely that:

Firstly, that not only the phoneme and articulation as such may be unfamiliar, but also the grapheme-to-phoneme conversion, and this may lead to unexpected pronunciation errors. Depending on the learners L1, it may hence be necessary to give additional feedback on the relation between the spelling and the expected pronunciation.

Secondly, that even if some pronunciation errors will be more probable and more frequent because of the relation between the L1 and the L2, other errors may also occur, and this must be taken into account in both the error detection and the feedback instructions. In particular, following articulatory instructions, the learners may produce sounds that are not easily classified as phonemes of the L1, nor of the L2.

Thirdly, the place of articulation and the articulator involved may influence how successful the learners are in following articulatory instruction. It would be of interest to carry out a more systematic experiment on articulation changes between isolated phonemes to explore different subjects' ability to consciously control different parts of the tongue. That is, to instruct them to start from a given articulation, e.g. [ə], and move the tongue in a specified manner, e.g. horizontally backwards, vertically upwards, diagonally upwards forward or backward, and monitor their success in following the instructions with ultrasound. We have postulated that many subjects will have difficulties following such instructions and have hence based much of our articulatory feedback on comparisons with words or sounds that are familiar from the L1 and have a similar articulation. A systematic exploration of the subjects' ability to follow general articulation instructions could serve as an important reference for improved articulation feedback.

A larger study, over several training sessions and using pre- and post-test listener ratings, is required to prove the long-term efficiency and retention of articulatory instructions, but the short-term changes in articulation observed in the current user test are nevertheless a positive indication that articulatory feedback instructions may be suitable in computer-assisted pronunciation training. This article has presented work in progress and we will continue to investigate how to present articulatory feedback effectively to the learners and how they respond to it.

Acknowledgments

This work is supported by the Swedish Research Council project 80449001 Computer-Animated LAnguage TEAchers (CALATEA).

Biography

Olov Engwall (born 1972) is an Associate Professor at KTH (Royal Institute of Technology) in Stockholm, Sweden. He has a MSc degree in engineering physics from KTH and has studied phonetics at Stockholm University. His PhD thesis focused on 3D articulatory modeling of the vocal tract, based on measurements using Magnetic Resonance Imaging, Electromagnetic articulography and Electropalatography. Since gaining his PhD, he has focused on applications using the 3D vocal tract model in an augmented reality talking head setting, such as supporting speech perception in noise and, in particular, investigating the use of embodied conversational agents in second language learning and pronunciation training. He has also

conducted several studies on acoustic-to-articulatory inversion.

References

Ananthakrishnan, G. and Engwall, O. (2011). Mapping between acoustic and articulatory gestures. *Speech Communication*, 53(4):567–589.

Ananthakrishnan, G., Neiberg, D., and Engwall, O. (2009). In search of non-uniqueness in the acoustic-to-articulatory mapping. In *Proceedings of Interspeech*, pages 2799–2802.

Bailly, G. and Badin, P. (2002). Seeing the tongue from outside. In *International Conference on Spoken Language Processing*, pages 1913–1916.

Bannert, R. (1994). *På väg mot svenskt uttal*. Studentlitteratur.

Beskow, J., Engwall, O., and Granström, B. (2003). Resynthesis of facial and intraoral motion from simultaneous measurements. In *International Congress of Phonetical Sciences*, pages 431–434.

Branderud, P. (1985). Movetrack – a movement tracking system. In *the French-Swedish Symposium on Speech*, pages 113–122.

Calinon, S., Guenter, F., and Billard, A. (2007). On learning, representing and generalizing a task in a humanoid robot. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 37(2):286–298.

Carroll, S. and Swain, M. (1993). Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition*, 15:357386.

Dent, H., Gibbon, F., and Hardcastle, W. (1995). The application of electropalatography (EPG) to the remediation of speech disorders in school-aged children and young adults. *International Journal of Language & Communication Disorders*, 30(2):264–277.

Deroo, O., Ris, C., Gielen, S., and Vanparys, J. (2000). Automatic detection of mispronounced phonemes for language learning tools. In *International Conference on Spoken Language Processing*, volume 1, pages 681–684.

Engwall, O. and Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers. *Computer Assisted Language Learning*, 20(3):235–262.

Engwall, O., Bälter, O., Öster, A.-M., and Kjellström, H. (2006). Designing the human-machine interface of the computer-based speech training system ARTUR based on early user tests. *Behavior and Information Technology*, 25:353–365.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, 51.

Fagel, S. and Madany, K. (2008). A 3-D virtual head as a tool for speech therapy for children. In *Interspeech*, pages 2643–2646.

Frankel, J., Wester, M., and King, S. (2007). Articulatory feature recognition using dynamic bayesian networks. *Computer, Speech and Language*, 21(4):620–620.

Gick, B., Bernhardt, M., Bacsfalvi, P., and Wilson, I. (2008). Ultrasound imaging applications in second language acquisition. In Edwards, J. and Zampini, M., editors, *Phonology and second language acquisition*, chapter 11, pages 309–322. John Benjamins Publishing Company.

Glisan, G., Dudt, K., and Howe, M. (1998). Teaching spanish through distance education: Implications of a pilot study. *Foreign Language Annals*, 31:48–66.

Hiroya, S. and Honda, M. (2004). Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Transactions on Speech and Audio Processing*, 12(2):175 – 185.

Jiang, J., Alwan, J., Keating, P., Auer, E., and Bernstein, L. (2002). On the

relationship between face movements, tongue movements, and speech acoustics. *EURASIP Journal on Applied Signal Processing*, 11:1174–1188.

Johnson, L. and Valente, A. (2008). Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. In *Conference On Innovative Applications Of Artificial Intelligence*, pages 1632–1639.

Johnson, W., Rickel, J., and Lester, J. (2000). Animated pedagocial agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, pages 47–78.

Kass, M., Witkin, A., and Tersopoulos, D. (1987). Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331.

Katsamanis, A., Papandreou, G., and Maragos, P. (2009). Face active appearance modeling and speech acoustic information to recover articulation. *IEEE Transactions on Audio, Speech and Language Processing*, 17(3):411–422.

Kjellström, H. and Engwall, O. (2009). Audiovisual-to-articulatory inversion. *Speech Commun.*, 51(3):195–209.

Kröger, B., Graf-Borttscheller, V., and Lowit, A. (2008). Two- and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders. In *Interspeech*, pages 2639–2642.

Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B., and Bhogal, R. (1997). The persona effect: Affective impact of animated pedagogical agents. In *Human Factors in Computing Systems*, pages 359–366.

Li, M., Kambhamettu, C., and Stone, M. (2009). Automatic contour tracking in ultrasound images. *Clinical Linguistics and Phonetics*, 19(6-7):545–554.

Lyster, R. (1998). Recasts, repetition, and ambiguity in l2 classroom discourse. *Studies in Second Language Acquisition*, 20:51–81.

Massaro, D., Bigler, S., Chen, T., Perlman, M., and Ouni, S. (2008). Pronunciation training: The role of eye and ear. In *Interspeech*, pages 2623–2626.

Massaro, D. and Bosseler, A. (2006). Read my lips: The importance of the face in a computer-animated tutor for autistic children learning language. *Autism: The International Journal of Research and Practice*, 10:495–510.

Massaro, D. and Light, J. (2003). Read my tongue movements: Bimodal learning to perceive and produce non-native speech /r/ and /l/. In *Eurospeech*, pages 2249–2252.

Massaro, D. and Light, J. (2004). Using visible speech for training perception and production of speech for hard of hearing individuals. *Journal of Speech, Language, and Hearing Research*, 47:304–320.

Modha, G., Bernhardt, M., Church, R., and Bacsfalvi, P. (2008). Case study using ultrasound to treat /ɹ/. *International Journal of Language & Communication Disorders*, 43(3):323–329.

Morton, H. and Jack, M. (2005). Scenario-based spoken interaction with virtual agents. *Computer Assisted Language Learning*, 18(3):171–191.

Neri, A., Cucchiarini, C., and Strik, H. (2006). Selecting segmental errors in L2 dutch for optimal pronunciation training. *International Review of Applied Linguistics*, 44:357–404.

Ouni, S. and Laprie, Y. (2002). Introduction of constraints in an acoustic-to-articulatory inversion. In *International Conference on Spoken Language Processing*, pages 2301–2304.

Richmond, K. (2006). A trajectory mixture density network for the acoustic-articulatory inversion mapping. In *Interspeech*, pages 577–580.

Scobbie, J., Wrench, A., and van der Linden, M. (2008). Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement. In *International Seminar on Speech Production*, pages 373–376.

Sheen, Y. (2004). Corrective feedback and learner uptake in communicative classrooms across instructional settings. *Language Teaching Research*, 8:263–300.

Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. In *Fonetik 2003*, pages 93–96.

Stepp-Greany, J. (2002). Student perceptions on language learning in a technological environment: Implications for the new millennium. *Language Learning & Technology*, 6(1):165–180.

Stone, M. and Davis, E. (1995). A head and transducer support system for making ultrasound images of tongue/jaw movement. *Journal of The Acoustical Society of America*, 98(6):3107–3112.

Strik, H., Truong, K., de Wet, F., and Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication*, 51(10):845–852.

Sugamura, N. and Itakura, F. (1986). Speech analysis and synthesis methods developed at ECL in NTT. *Speech Commun.*, 5:199–215.

Teppermann, J. and Narayanan, S. (2008). Using articulatory representations to detect segmental errors in nonnative pronunciation. *IEEE Transactions on Audio, Speech and Language Processing*, 16(1):8–22.

Toda, T., Black, A., and Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian Mixture Model. *Speech Communication*, 50(3):215–227.

van Mulken, S., André, E., and Müller, J. (1998). The persona effect: How substantial is it? In *Human-Computer Interaction*, pages 53–66.

Wei, S., Hu, G., Hu, Y., and Wang, R.-H. (2009). A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication*, 51(10):896–905.

Wik, P. and Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech Communication*, 51(10):1024–1037.

Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behaviour. *Speech Communication*, 26:23–43.