PAPER 8.

Beskow, J. (submitted4). Trainable Articulatory Control Models for Visual Speech Synthesis, submitted to *International Journal of Speech Technology*.

TRAINABLE ARTICULATORY CONTROL MODELS FOR VISUAL SPEECH SYNTHESIS

Jonas Beskow

KTH, Centre for Speech Technology, SE-10044 Stockholm, Sweden

ABSTRACT

This paper deals with the problem of modelling the dynamics of articulation for a parameterised talking head based on phonetic input. Four different models are implemented and trained to reproduce the articulatory patterns of a real speaker, based on a corpus of optical measurements. Two of the models are based on coarticulation models from speech production theory and two are based on artificial neural networks, one of which is specially intended for streaming real-time applications. The different models are evaluated through comparison between predicted and measured trajectories, as well as through a perceptual intelligibility experiment. Results show that all models give significantly increased speech intelligibility over the audio-alone case, but none of the models can be said to outperform the others.

Keywords: Speech synthesis, facial animation, coarticulation, artificial neural networks, perceptual evaluation

1 INTRODUCTION

During the last decade, there has been an increasing interest in computer animated talking heads. The applications for this technology include virtual language tutors (Cole et al., 1998), communication aids for hard-of-hearing people (Agelfors et al., 1998), embodied conversational agents in spoken dialogue systems (Gustafson et al. 2000) and talking computer game characters, to name only a few. Proper visual speech movements are often a crucial factor for the realism of such avatars, and in some applications, realistic speech movement constitute the main motivation for the technology (Agelfors et al, 1998; Siciliano et al., 2003). It is well established that a view of the speaker's face leads to increased speech intelligibility. This has been shown for natural speech (Sumby & Pollack, 1954) as well as for synthetic visual speech in several languages (LeGoff et al., 1994; Massaro 1998; Siciliano et al., 2003).

Visual speech synthesis can be accomplished either through manipulation of video images (Bregler et al., 1997; Brooke and Scott, 1998; Ezzat et al., 2002) or based on two- or three dimensional models of the human face and/or speech organs that are under control of a set of deformation parameters (Beskow, 1997; Cohen & Massaro, 1993; Pelachaud et al., 1996; Pelachaud, 2002; Reveret et al., 2000).

In most implementations we can make a distinction between the visual signal model and the control model. The visual signal model is responsible for producing the image, given a vector of control parameter values. The control model is responsible for driving the animation by providing these vectors to the signal model at each point in time, given some symbolic specification of the desired animation. In a general sense, the input to a control model could contain information about speech articulation, emotional state, discourse information, turntaking etc. In this study we will restrict ourselves to studying *articulatory* control models. An articulatory control model can be described as a process that produces control parameter trajectories to govern articulatory movements for a given phonetic target

specification, typically a sequence of time labelled phonemes, optionally including stress and phrasing markers. This is normally the kind of information that is available at the phonetic stage of a text-to-speech system. One problem an articulatory control model has to deal with is that of coarticulation.

1.1 Coarticulation in speech production theory

Coarticulation refers to the way in which the realisation of a phonetic segment is influenced by neighbouring segments. It is the result of articulatory planning, inertia in the biomechanical structures of the vocal tract, and economy of production. But coarticulation also serves a communicative purpose in making the speech signal more robust to noise by introducing redundancies, since the phonetic information is spread out over time. Two types of coarticulation can be identified: backward and forward coarticulation.

Backward, or carry-over coarticulation, refers to the way in which articulation at some point in time is affected by the articulation at some previous point in time. This effect has sometimes been attributed to inertia in the motor system, but it has also been shown to be under deliberate neurological control (MacNeilage & DeClerk, 1969).

Forward, or anticipatory coarticulation, on the other hand, is a term used to describe how articulation at some point in time is affected by articulation of segments not yet realised. This effect cannot be explained by biomechanical properties, but rather there must be a higher level of articulatory planning involved.

Explaining the variation that occurs in human speech production as a result of coarticulation is a fundamentally difficult problem that has been the subject of many studies (Bladon and Al-Bamerni, 1976; Löfqvist, 1990; MacNeilage & DeClerk, 1969; Perkell, 1990; Öhman, 1967). Following Perkell (1990), existing models of coarticulation can be divided into look-ahead models and time-locked models. In look-ahead models an anticipatory coarticulatory gesture begins at the earliest possible time allowed by the articulatory constrains of other segments in the utterance. In a V_1CV_2 utterance where V_1 is unrounded and V_2 is rounded, onset of the rounding gesture begins at the offset of V_1. In Öhman's (1967) model, the time varying shape of the vocal tract is modelled as a vowel gesture where the vocal tract gradually changes shape from V_1 to V_2, onto which a consonant gesture is superimposed. The consonant has an associated temporal blend function that dictates how its shape should blend with the vowel gesture over time. It also has a spatial coarticulation function that dictates to what degree different parts of the vocal tract should deviate from the underlying vowel shape, i.e. how context-sensitive the realisation of that segment is for the different articulators. Bladon and Al-Bamerni (1976) suggest the term *coarticulation resistance* to indicate to what extent a particular segment is susceptible to coarticulation; a segment with high coarticulation resistance will have more or less the same realisation regardless of context, whereas a segment with low coarticulation resistance will realised in a highly context-dependent way.

In a time-locked model the onset of a gesture occurs a fixed time before the onset of the associated segment, regardless of the timing of other segments in the utterance. In Löfqvist's gestural model (Löfqvist, 1990), speech production is modelled as a series of overlapping articulatory gestures. Each segment has a temporal dominance function that gradually increases up to a peak value and then decreases. The dominance functions of the different segments dictate the temporal blending of articulatory commands related to these segments. The height of the dominance function at the peak determines to what degree the segment is subject to coarticulation, i.e. the coarticulation resistance of the articulator for that segment.

Perkell (1990) also suggests a hybrid between the look-ahead and time-locked models, where the gesture for V_2 is divided into a gradual initial phase which starts at the offset of V_1 and a more rapid second phase that begins at a fixed time before V_2.

1.2 Coarticulation in multimodal speech synthesis

Early text-to-speech systems (Allen et al., 1987; Carlson et al., 1982) employed parametrically controlled models for speech generation. In these systems, control parameter trajectories were generated within a rule-based framework, where coarticulatory effects were modelled using explicit rules. Later, as concatenative speech synthesis techniques grew popular, the need for such rules diminished, since coarticulation was inherently present in the speech units used, for example diphones, demi-syllables or the arbitrary sized units used in contemporary unit-selection based speech synthesisers.

Rule based control schemes have been successfully employed for visual speech synthesis. Pelachaud et al. (1996) describe an implementation of the look-ahead model. Phonemes are clustered into visemes that are classified with different deformability rank, which serves to indicate to what degree that viseme should be influenced by its context (c.f. coarticulation resistance in the previous section). Visemes with low deformability serve as key-shapes that influence the shape of the more deformable ones.

Another rule-based look-ahead model is proposed by Beskow (1995). In this model, each phoneme is assigned a target vector of articulatory control parameters. To allow the targets to be influenced by coarticulation, the target vector may be under-specified, i.e. some parameter values can be left undefined. If a target is left undefined, the value is inferred from context using interpolation, followed by smoothing of the resulting trajectory. As an example, consider the lip rounding parameter in a V_1CCCV_2 utterance where V_1 is unrounded and V_2 is rounded. Lip rounding would be unspecified for the consonants, leaving these targets to be determined from the vowel context by linear interpolation from the unrounded V_1, across the consonant cluster, to the rounded V_2.

For visual speech synthesis, approaches based on concatenation of context dependent units have been less dominant, although they have been used for video based synthesis (Bregler et al., 1997) as well as in model based systems (Hällgren and Lyberg, 1998).

The model described by Cohen and Massaro (1993) is based on Löfqvist's (1990) gestural theory of speech production. In this model, each segment is assigned a target vector. Overlapping temporal dominance functions are used to blend the target values over time. The dominance functions take the shape of a pair of negative exponential functions, one rising and one falling. The height of the peak and the rate with which the dominance rises and falls are free parameters that can be adjusted for each phoneme and articulatory control parameter. Because the rise and fall times are context-independent, the Cohen-Massaro model can essentially be regarded as a time-locked model. It should be noted however that due to the nature of negative exponential functions, all dominance functions extend to infinity, so in practice onset of a gesture occurs gradually as the dominance of the gesture rises above the dominance of other segments. The free parameters were empirically determined through hand-tuning and repeated comparisons between synthesis and video recordings of a human speaker.

Other investigators have proposed enhancements to the Cohen-Massaro model as well as data-driven automatic estimation of its parameters. Le Goff (1997) modified the dominance functions to be n-continuous and used trajectories extracted from video recordings of a real speaker uttering $V_1CV_2CV_1$ words to tune the model for French. The resulting model was perceptually evaluated using an audiovisual phoneme identification task with varying levels of background acoustic noise.

The Cohen-Massaro model offers no way to ensure that certain targets are reached, such as the closure in a bilabial stop, which can be problematic especially for short segments. To overcome this problem, Cosi et al. (2002) augment the Cohen-Massaro model with a resistance function that can be used to suppress the dominance of surrounding segments, thereby forcing the attainment of the target. Parameters of the augmented model were

estimated from a database of symmetrical VCV utterances recorded using an optical motion tracking system.

Recently, Massaro et al. (in press) used optical motion tracking of 19 points on a real speaker's face to obtain a database of 100 sentences that were used to tune the free parameters of the original Cohen-Massaro model. Training was carried out both for a set of 39 monophones as well as for 509 context-dependent phones.

Reveret et al. (2000) adopt Öhman's model of coarticulation to drive a French speaking talking head. Coarticulation coefficients and the temporal functions guiding the blending of consonants and the underlying vowel track were estimated from a corpus derived using video analysis of 24 VCV words.

A different type of control model is proposed by Pelachaud (2002). This model uses radial basis functions (RBFs) to model the trajectories of articulatory parameters. The RBFs share some properties with the dominance functions of the Cohen-Massaro model, in that they are negative exponential functions, but in Pelachaud's model three RBFs per segment and parameter are used as opposed to one single dominance function. The free parameters of the RBFs were estimated from a training corpus of optically tracked VCV-sequences.

1.3 Comparing the models

A number of factors make it difficult to assess the relative merits of the different approaches mentioned above. The corpora on which the models are trained differ in size, type of content (VCV words, sentences etc.) as well as language. Different control parameter sets are used, obtained using a variety of techniques, and different facial models are used to produce the resulting animations.

The purpose of the present study is to seek an answer to the question whether there is reason to prefer one of these models to another when building a data-driven visual speech synthesis system. The two main classes of models, look-ahead (represented by Öhman's model) and time-locked (represented by Cohen & Massaro's model) are trained on a corpus of phonetically rich sentences and then evaluated objectively as well as perceptually, on a test set not part of the original training data. In addition, they are compared against two variants of a novel control model based on recurrent time delayed artificial neural networks, one of which is designed specifically for real time applications.

1.4 Real-time considerations

All of the models described above account for anticipatory coarticulation, which is indeed a fundamental property of human speech production. But any control model attempting to model anticipation must have information about upcoming segments ahead of time. In the case of look-ahead models, the amount of time needed is dependent on the timing and identity of the particular segments involved. In practice, many implementations will require access to the full utterance before any trajectories can be computed. This is for example the case with an unconstrained implementation of the Cohen-Massaro model, where the negative exponential dominance functions extend to infinity.

In certain real world applications, such models are impossible to use without modifications. In the Synface project (Siciliano et al., 2003) the goal is to develop a communication aid for hard of hearing people, consisting of a talking head faithfully recreating the lip movements of a speaker at the other end of a telephone connection in (close to) realtime, based only on the acoustic signal. The system will utilise a phoneme recogniser, outputting a stream of phonemes that should be instantly converted into facial motion. In this scenario there is no way to know what phonemes will arrive in the future, so any model attempting anticipatory coarticulation will fail. To allow for this kind of application, we need control models that can do a reasonable job given a very limited look-

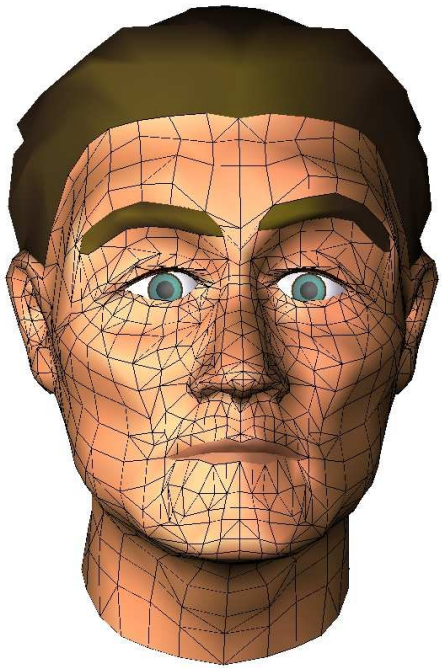

Figure 1. The parametrically controlled talking head.

ahead window (typically less than 50 milliseconds). In the present study, one such low-latency model is implemented and compared against the other models.

2 TALKING HEAD MODEL

The geometrical talking head model used in this study is a parametrically controlled deformable polygon surface, depicted in figure 1, based on the work described in Beskow (1995, 1997). Articulatory deformations are implemented as weighted geometrical transforms (translations, rotations and scalings) that are applied to the vertices of the polygon mesh, according to principles first introduced by Parke (1982). In this study, data-driven articulation is controlled by the ten parameters *jaw rotation, lip rounding, upper lip retraction, upper lip raise, lower lip retraction, lower lip depression, left mouth corner stretch, right mouth corner stretch, jaw thrust* and *jaw shift* (sideways).

3 DATA ACQUISITION AND PROCESSING

Audio and facial motion was recorded for 287 phonetically rich Swedish sentences. The sentences were extracted from a large corpus of newspaper text to optimise coverage of tri-phones.

Facial motion was recorded using an optical motion tracking system from Qualisys (`http://www.qualisys.se`) with four cameras. The system tracks passive reflective markers (4 mm in diameter) and calculates their 3D-coordinates at a rate of 60 frames per second.

A native Swedish male non-professional speaker was instrumented with 30 markers, glued to the speaker's jaw, cheeks, lips, nose and eyebrows. In addition, five markers were affixed to a spectacle frame worn by the speaker to establish a reference for global head movement.

Audio was recorded onto DAT-tape. A synchronisation signal from the Qualisys system consisting of a pulse train with one pulse per captured frame was recorded on one channel of the DAT to facilitate post-synchronisation of audio and motion data.

3.1 Phonetic labelling

The audio signal was used to phonetically label the corpus. First, the sentences were phonetically transcribed using the transcription part of a text-to-speech system. The transcriptions were manually corrected to match the actual pronunciations and stress patterns observed in the recorded utterances. A viterbi-based forced alignment procedure, with speaker adaptation of the acoustic models, was used to obtain the temporal alignment of the phonetic labels (Young et al. 1997). The symbol set included the long vowels [ɑ:, o:, ɑ:, i:, e:, ɛ:, y:, ø:, ʉ:], the short vowels [ʊ, ɔ, a, ɪ, e, ɛ, æ, ʏ, œ, œ̞, ø, ɵ, ə] and the consonants [p, t, ʈ, k, b, d, ɖ, g, m, n, ɳ, ŋ, l, ɭ, f, s, ʃ, ç, j, r, v, h]. In addition, a symbol for silence and six symbols denoting release phase of stops were used. To encode stress information, an additional set of 23 symbols for stressed vowels were included, resulting in a total of 76 symbols.

3.2 Pre-processing of movement data

Since we are interested only in the articulatory movements in this study, the first step in the processing of the motion data was to factor out global translations and rotations of the head. Using the reference markers on the spectacle frame, which can be considered fixed with respect to the head, a head local coordinate system was defined, onto which the coordinates of all other points were transformed. This transform was applied to each frame of movement data.

3.3 Fitting the model to data

To be able to use the recorded motion data to control our models, we need a way to map the motion of the 30 points into control parameter trajectories for the face model. This process is non-trivial for several reasons. Firstly, many of the face model parameters affect the same regions of the face. For example, the points on the lower lip are affected by lip raising as well as jaw opening. Thus, a global optimisation is needed to find the best possible combination of parameter values for a given frame of motion data. Another factor we need to take into account is that the geometry of the face model may not match the recorded speaker. In fact, we want the process to be general enough to do a reasonable job of mapping the motion data to any human-like face, with different proportions than those of the recorded speaker. Since recording of 3D-motion data is a complicated task, it is beneficial if we can reuse the data with different facial models.

A general solution to the problem is to use multi-dimensional minimisation techniques, to fit the model to the recorded data set by adjusting its parameters. This reduces the problem to that of defining a proper error metric for a match between model and data. The model fitting is divided into two stages. First a static orientation fit, which applies a linear transform to the measurement data that ensures optimal spatial alignment for a particular reference frame. (It should be noted that for practical reasons the static fit actually fits the data to the model rather than the opposite, but since it is a linear transform this is equivalent.) Secondly a dynamic shape fit to find optimal facial deformation parameters is carried out for every frame in the data set. For these minimisations, Powell's direction set method is used (Press et al., 1992), which is considered a robust general-purpose method when the gradient of the error function is unknown.

3.4 Orientation fit

As stated before, we cannot expect a perfect match between measurements and model, since the model is not based on the measured subject. Nevertheless, we want the important features of the two faces to align as closely as possible. Since we have already compensated for global rotation and translation, we try to find a static transform $\mathbf{T}(\mathbf{x})$ that gives the best match between data and model for a certain reference frame. The transform $\mathbf{T}(\mathbf{x})$ is composed of matrices for translation in three dimensions, rotation around three axes and scaling, yielding a total of seven free parameters $\mathbf{x} = (t_x, t_y, t_z, \alpha_x, \alpha_y, \alpha_z, s)$.

A number of strategic points were selected at mouth corners, tip of the nose, chin and eyebrows, where a corresponding model vertex $\mathbf{v}_i$ for each marker position $\mathbf{p}_i$ was identified. The error function

$$e(\mathbf{x}) = \sum_{i \in K} \left| \mathbf{v}_i - \mathbf{p}_i \cdot \mathbf{T}(\mathbf{x}) \right| \quad (1)$$

where K is the set of strategic points, was minimised with respect to $\mathbf{x}$, and the resulting best-fitting transform was applied to all frames in the data set.

3.5 Shape fit

For the dynamic shape fit, the goal is to find an optimal vector of model parameters $\mathbf{y} = (y_1, y_2 \ldots y_M)$ given a frame of measurement points $\mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_N$. Minimisation will help us find this optimal vector, but what is an adequate error metric to use for this minimisation? One possibility is to extend the error function that was used in the orientation fit to include all markers, by pairing each marker with the model vertex closest to it in the reference frame. The problem is that this error metric will penalise inherent static differences in geometry, and try to compensate for them using the articulatory parameters, which will lead to inaccurate articulation. To circumvent this problem, an error metric was defined in such a way that it equals zero when both the measured face and the virtual face are in a well-defined reference pose. Rather than matching the markers directly against existing model vertices, a set of *virtual markers* was introduced on the model, that are tied to a region of the model, and move with it, but need not coincide with existing vertices or even lie on the surface of the model. A specially recorded reference pose, with relaxed face and closed jaw and lips, was used to tie the virtual markers to the face model, which was manually adjusted to match this reference pose.

Mathematically, a virtual marker $\mathbf{q}$ is defined in terms of *three* non co-linear model vertices, $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$. The three vertices together with the surface normal vector, form a coordinate system in which we can express the position of $\mathbf{q}$ in terms of the coordinates s, t and u:

$$\mathbf{q} = \mathbf{v}_1 + s(\mathbf{v}_2 - \mathbf{v}_1) + t(\mathbf{v}_3 - \mathbf{v}_1) + u\mathbf{e}_1 \quad (2)$$

where $\mathbf{e}_1$ is the normalized vertex normal of vertex $\mathbf{v}_1$. The s and t coordinates represent $\mathbf{q}$'s projection in the plane of ($\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$) in barycentric coordinates, i.e. along the edges ($\mathbf{v}_1$, $\mathbf{v}_2$) and ($\mathbf{v}_1$, $\mathbf{v}_3$) respectively. The u coordinate represents offset from the surface in the direction of the vertex normal.

For each real, measured marker $\mathbf{p}_i$, a corresponding virtual marker $\mathbf{q}_i(\mathbf{y})$ was defined, where $\mathbf{y}$ is the vector of articulatory parameters of the model (listed in section 2). Thus we can define the following error metric:

$$e(\mathbf{y}) = \sum_{i=1}^{N} \left| \mathbf{q}_i(\mathbf{y}) - \mathbf{p}_i \right| \quad (3)$$

The minimisation to find $\mathbf{y}_{\min}$ was carried out for each frame of motion data in the corpus. The result is a matrix $\mathbf{Y}$ where each row consists of $\mathbf{y}_{\min}$ for that frame. This matrix contains a re-synthesis of the recording, i.e. a set of parameter trajectories that can be used to

animate the face model to follow the recorded movement data as closely as possible. This matrix constitutes the training and test data for the articulatory control models described in the following section.

4 MODELS AND TRAINING

Given the articulatory parameter trajectories estimated in the preceding section, and the time-aligned phonetic labels, we want to train different articulatory control models, to see which one is best capable of reproducing the patterns observed in the data. In the previous section we were fitting the facial model to a set of 3D points, now we are fitting articulatory models to a set of parameter tracks. In both cases we use the same general-purpose technique: minimisation of global error.

The models that are being evaluated vary quite a lot in the ways in which they predict articulatory trajectories. What is true for all models is that they have a certain number of free parameters represented as a vector **x**, take a phonetic input specification, and output a set of estimated articulatory trajectories **Z**. The input consists of a time-labelled phone sequence with stress information. As stated in section 3.1, stress information was encoded by separate symbols for stressed vowels, which effectively doubles the number of vowel symbols. A larger symbol set leads to fewer examples of each symbol in the data set, which potentially can lead to worse generalisation by the models. To evaluate this risk, all models were trained with and without stress information, and they all performed slightly better with the larger, stress-inclusive symbol set.

The output of a control model is a set of predicted control parameter trajectories that can be compactly represented as a matrix **Z** where each row is a time frame and each column is a control parameter. The predicted trajectories are compared with the target trajectories **Y**. The objective of training the control models is to make **Z** and **Y** as similar as possible by varying the free parameters **x**. If we let $\mathbf{z}_i$ and $\mathbf{y}_i$ denote the i:th column of **Z** and **Y** respectively and N is the number of articulatory parameters, then the error to minimize can be expressed as

$$e(\mathbf{x}) = \sum_{i=1}^{N} (\mathbf{z}_i - \mathbf{y}_i)^{\mathrm{T}} (\mathbf{z}_i - \mathbf{y}_i) \tag{4}$$

or, if we model each articulatory parameter individually

$$e(\mathbf{x}) = (\mathbf{z} - \mathbf{y})^{\mathrm{T}} (\mathbf{z} - \mathbf{y}) \tag{5}$$

When minimising functions with a large number of free parameters, such as the control models investigated here, taking advantage of gradient information can greatly reduce the amount of computation required to find the minimum of the function. The gradient of the error function is obtained by taking the derivative of (5) with respect to each of the components of **x**:

$$\nabla e(\mathbf{x}) = \left(\frac{\partial e(\mathbf{x})}{\partial x_0}, \frac{\partial e(\mathbf{x})}{\partial x_1} \ldots \frac{\partial e(\mathbf{x})}{\partial x_N} \right) \quad \text{with} \quad \frac{\partial e(\mathbf{x})}{\partial x_k} = 2 \left(\frac{\partial \mathbf{z}}{\partial x_k} \right)^T (\mathbf{z} - \mathbf{y}) \tag{6}$$

Thus, if we can calculate the derivative of **z** with respect to each of the components of **x**, then we can also calculate the gradient of $e(\mathbf{x})$.

4.1 Training procedure

The 287 sentence corpus was split into a training part consisting of 200 sentences and a test part of 87 sentences. After each iteration of training, the global error on the test set was computed. To prevent over-training, training was terminated when the error on the test set stopped decreasing.

4.2 Cohen and Massaro's model of coarticulation

In the Cohen-Massaro model, each articulatory parameter trajectory is described as a weighted sum of target values for all segments of the utterance. The weight of a target changes over time according to the dominance function associated with that segment. The dominance function has a peak at the centre of the segment, and an exponential rise and fall-off to the sides. The dominance functions overlap, which makes all segments in the utterance influence each other. The dominance function for segment i is given by

$$D_i(t) = \begin{cases} \alpha_i \cdot e^{-\theta_i(\tau_i - t)^{c_i}} & t < \tau_i \\ \alpha_i \cdot e^{-\varphi_i(t - \tau_i)^{c_i}} & t \geq \tau_i \end{cases} \tag{7}$$

where τ_i is the centre time of the segment. a_i is a scaling factor used to control the degree of dominance for each segment. θ_i and φ_i are coefficients for forward- and backward coarticulation respectively. Lower values for these coefficients will cause the dominance function to spread out further into following or preceding segments respectively. The exponent c_i is used to vary the shape of the dominance function. A lower value produces a function with a sharp peak, whereas a higher value produces a more rounded peak.

Given the dominance function $D_i(t)$ and the target value T_i for each segment, the value of parameter z at the time t is given by

$$z(t) = \frac{\sum_{i=1}^{N} T_i D_i(t)}{\sum_{i=1}^{N} D_i(t)} \tag{8}$$

Where N is the number of segments in the utterance. Dominance functions, targets and synthesised trajectory for a three-segment utterance can be seen in figure 2. The segment model has five free parameters: T, α, ϕ, θ and c. Given 76 phones and 10 articulatory parameters, this yields a total of 3800 free parameters to be estimated from the data set. But since each articulatory parameter is modelled in isolation, we can treat it as ten separate estimation problems of 380 parameters each.

Regarding c as a free parameter that can be set individually for each phoneme is a slight

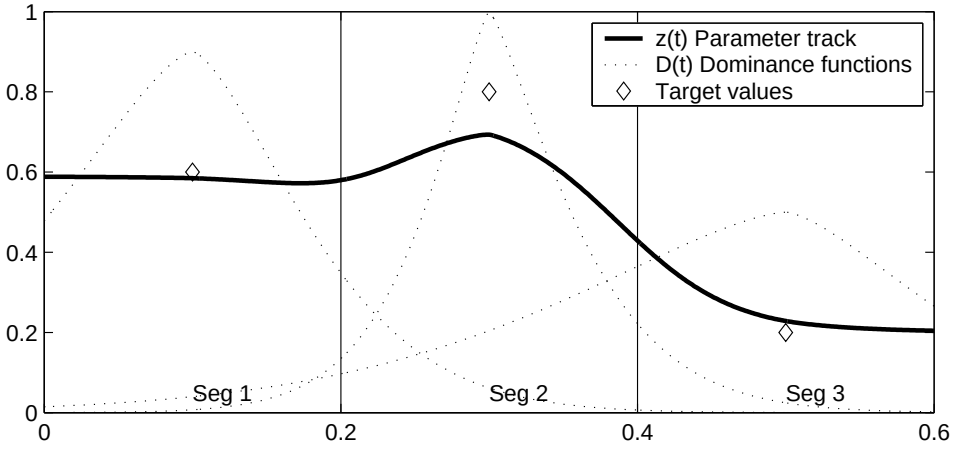

Figure 2. Cohen-Massaro model of a three-segment utterance.

alteration to the original Cohen-Massaro model, in which c was regarded as a global constant.

The gradient equation (6) requires that we compute the derivative of $z(t)$ with respect to each of the free parameters. Differentiating (8) w.r.t. T_i yields

$$\frac{\partial z(t)}{\partial T_i} = \frac{D_i(t)}{\sum_{j=1}^{N} D_j(t)} \tag{9}$$

For α_i, φ_i, θ_i and c_i we can use the chain-rule and differentiation of quotient to obtain the general form

$$\frac{\partial z(t)}{\partial \psi_i} = \frac{\partial D_i(t)}{\partial \psi_i} \cdot \frac{T_i \sum_{j=1}^{N} D_j(t) - \sum_{j=1}^{N} T_j D_j(t)}{\left(\sum_{j=1}^{N} D_j(t) \right)^2} \tag{10}$$

where ψ_i can be substituted with one of α_i, θ_i, φ_i, or c_i.
For each of these parameters, the derivative of d_i is given by

$$\frac{\partial D_i(t)}{\partial \alpha_i} = \begin{cases} e^{-\theta_i(\tau_i - t)^{c_i}} & t < \tau_i \\ e^{-\varphi_i(t-\tau_i)^{c_i}} & t \geq \tau_i \end{cases} \tag{11}$$

$$\frac{\partial D_i(t)}{\partial \theta_i} = D_i(t) \cdot \begin{cases} -(\tau_i - t)^{c_i} & t < \tau_i \\ 0 & t \geq \tau_i \end{cases} \tag{12}$$

$$\frac{\partial D_i(t)}{\partial \varphi_i} = D_i(t) \cdot \begin{cases} 0 & t < \tau_i \\ -(t-\tau_i)^{c_i} & t \geq \tau_i \end{cases} \tag{13}$$

and

$$\frac{\partial D_i(t)}{\partial c_i} = D_i(t) \cdot \begin{cases} \theta_i(\tau_i - t)^{c_i} \ln(\tau_i - t) & t < \tau_i \\ \varphi_i(t-\tau_i)^{c_i} \ln(t-\tau_i) & t \geq \tau_i \end{cases} \tag{14}$$

Equations (9) and (11-14) substituted into (10) give the partial derivatives of the parameter trajectory $z(t)$ with respect to the free parameters of segment i. The gradient (6) contains derivatives of the free parameters for each phoneme class, rather than for each segment of the utterance. To obtain the gradient components associated with phoneme p, the derivative trajectories are summed over all i for which segment i equals p. The resulting trajectories are then substituted into (6) to obtain the components of the gradient of the error function.

The model was trained using the Matlab function `fminunc`, which implements the Gauss-Newton minimisation algorithm.

4.3 Öhman's model of coarticulation

Öhman (1967) describes a model of coarticulation where the trajectory of an articulator is modelled as a vowel track with superimposed consonantal gestures. The model was originally intended to describe lingual coarticulation in VCV sequences by predicting cross-sectional distances in the vocal tract. Following Reveret et al. (2000), we use the model to predict general articulatory parameters and extend it to cope with arbitrary phone sequences.

The vowel track $v(t)$ is formed by interpolation between successive vowel targets. A consonant is specified by a target value c, a coarticulation factor w_c and a function $k(t)$ that dictates the temporal blending of vowel track and consonant target, where w_c and $k(t)$ are in

the interval [0, 1]. According to Öhman, the trajectory of a given articulatory parameter over a VCV sequence can be described as

$$z(t) = v(t) + w_c k(t)\big(c - v(t)\big) \tag{15}$$

In order to apply Öhman's equation to arbitrary phoneme sequences, we can write

$$z(t) = v(t) + \sum_{i \in C} w_{ci} k_i(t)\big(c_i - v(t)\big) \tag{16}$$

where C is the set of all consonants in the utterance. Note that the Öhman model does not define coarticulation between consonants, so the influence of one consonant must extend no further than to the peak of the preceding or following gesture. If we let τ_n denote the centre of segment n, and assume that the peak of each gesture occurs at the centre of the segment, then a suitable blend function $k_i(t)$ for segment i should go from zero at $t = \tau_{i-1}$, to one at $t = \tau_i$ and back to zero at $t = \tau_{i+1}$. Furthermore, to make the resulting trajectory smooth with a continuous derivative, we want $k'(\tau_{i-1}) = k'(\tau_{i+1}) = 0$. We choose to represent $k(t)$ as a third degree polynomial that fulfils the above criteria.

The vowel track $v(t)$ is formed by temporally blending successive fixed vowel targets a_i, according to the function

$$v(t) = \frac{\sum_{j=1}^{N_V} a_j b_j(t)}{\sum_{j=1}^{N_V} b_j(t)} \tag{17}$$

where N_V is the number of vowels in the utterance and $b_j(t)$ is the blend function of the j:th vowel in the utterance. Since there is no intervocalic coarticulation in the Öhman model, the reasoning concerning the blend function for consonants in the previous paragraph applies to vowels as well, thus we can use the same polynomial for $b_j(t)$; a cubic function reaching one at the centre of vowel j and zero at the centre of the preceding (j-1) and following (j+1) vowel. An illustration of the functions involved can be seen in figure 3.

The free parameters in our implementation of the Öhman model are the vowel target values a_n, the consonant target values c_n, and the consonant coarticulation factors w_{c_n}. Given 48 vowel phones (24 unstressed + 24 stressed) and 28 consonants, we have 104 free parameters to estimate for each articulatory parameter to be predicted. In order to calculate the gradient of the error function to optimise the training process, we obtain the derivative

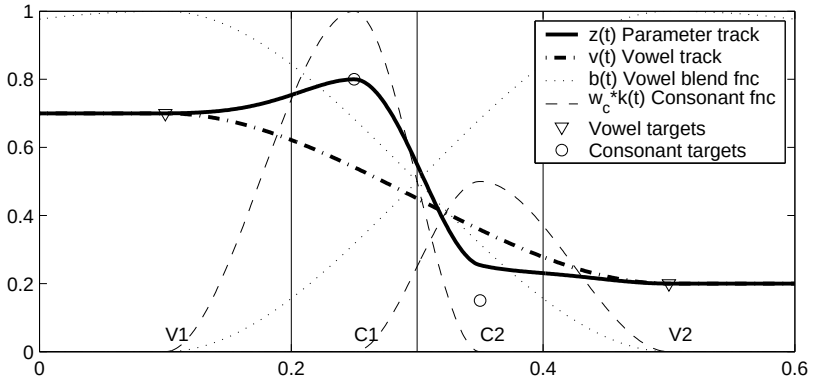

Figure 3. Öhman model of a $V_1C_1C_2V_2$ utterance.

of (16) with respect to each of the free parameters:

$$\frac{\partial z(t)}{\partial w_{ci}} = k_i(t)\big(c_i - v(t)\big) \tag{18}$$

$$\frac{\partial z(t)}{\partial c_i} = w_{ci} k_i(t) \tag{19}$$

and

$$\frac{\partial z(t)}{\partial a_i} = \frac{b_i(t)}{\sum_{j=1}^{N_V} b_j(t)} \left(1 - \sum_{i \in C} w_{ci} k_i(t) \right) \tag{20}$$

For each phoneme *p*, these trajectories are summed over all *i* where segment *i* equals *p*, and substituted into (6) to obtain the gradient of the error function.

The model was trained using the Gauss-Newton minimisation algorithm.

4.4 ANN 1 - symmetrical context

Artificial neural networks (ANNs) were trained to predict articulatory parameter values on a frame-by-frame basis. Input to the networks consisted of a series of feature vectors, one per time frame (at 60 frames per second) constructed from the time-labelled phone sequence using table lookup. The vectors contained 17 phonetic features uniquely describing the present phone. The features set consisted of *phoneme class* (consonant or vowel), six binary place features (*bilabial, labiodental, dental, retroflex, alveolar* and *velar*) five binary manner features (*nasal, fricative*, *stop*, *release* and *voiced*), four 3-valued vowel features (*open, front, rounded* and *protruded*) and *stress*.

A large number of network configurations were examined before the final configuration was chosen. For each of the variables below, several networks were trained and evaluated based on the error over the test set and subjective judgements of the quality of the resulting animations. Due to the combinatory complexity it was impossible to try all combinations of all variables.

- *Segment- vs. frame based input* – one input vector per segment (with an additional duration feature), versus synchronous frame based input. The latter led to better convergence and lower error in the estimation.
- *Single vs. multiple networks* – one single network with ten outputs vs. multiple networks, each predicting a subset of the parameters, vs. ten single-output networks predicting one parameter each.
- *Number of hidden units* – for all different configurations varying numbers of hidden units were used.
- *Size of context window* – varying number of frames of forward and backward context was evaluated.
- *Recurrent v.s feed-forward* – recurrent networks tended to produce smoother parameter trajectories than the feed-forward networks.

The best performance was achieved with a set of four separate recurrent three-layer ANNs, each predicting its own group of two or three parameters. It was found that by grouping the parameters involved in bilabial occlusion (jaw rotation, upper lip raise and lower lip depression) into a separate network, more reliable prediction of occlusion was achieved. Other groups were formed for rounding, protrusion and jaw motion, as shown in table 1. A time delay was used for the input to provide a symmetric context of ±15 frames, yielding a total input layer size of 17×31 = 527 units. The hidden layer had 30 to 50 units and was connected to itself with time delays between one and five frames. The networks had 21332

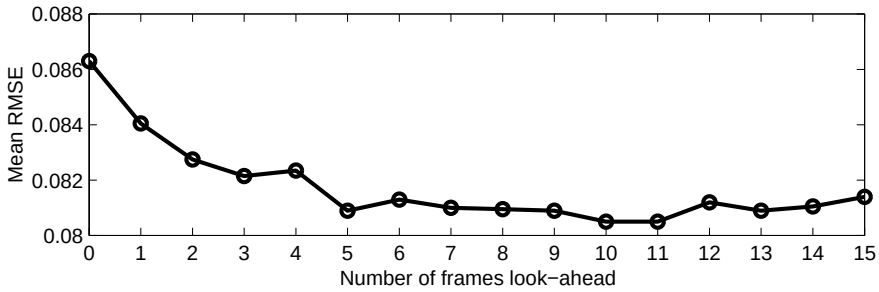

Figure 4. Total error of the 16 ANNs trained with varying degree of look-ahead, from 0 to 15 frames.

to 40603 connections (see table 1), with a total of 123870 free parameters for the four networks combined.

The networks were trained using the algorithm back propagation through time, as implemented by (Ström, 1997). Training was terminated when the error over the test set stopped decreasing, after approximately 200 iterations.

4.5 ANN 2 - low latency model

For the purpose of real-time applications, as discussed in section 1.4, a second group of networks was trained where the context was asymmetrically shifted to require less latency. In order to find a good compromise between low latency and high accuracy, a series of 16 networks was trained where the forward context varied from zero (no latency) to fifteen frames (250 milliseconds; each frame represents one 60th of a second). The networks were based on the occlusion ANN in the previous section, and the total context window size (look-ahead, look back and current) was 31 frames for all networks. All networks were trained for 200 iterations. The resulting error over the test set is shown in figure 4.

Based on this information, it was decided that two frames of look-ahead (33 milliseconds) provided a good compromise between accuracy and low latency. Four networks, identical to the ones trained in the previous section, with the exception of the context windows being 2 frames forward and 28 frames backward, were trained using the same procedure as above.

5 OBJECTIVE EVALUATION

To evaluate the performance of the models, target and predicted trajectories were compared for the four models, over the 89-sentence test set. Figure 5 shows an example of parameter traces for the jaw rotation parameter for each of the models for the test set utterance "*svenska kyrkan ska vara till hjälp för de församlingsbor som så önskar*" ("*the Swedish church should*

Table 1. The four ANNs and the parameter groups they predict

Network	Predicted parameters	Number of hidden units	Number of connections
Occlusion	Jaw rotation, upper lip raise, lower lip depression	50	40603
Rounding	Lip rounding, left mouth corner stretch, right mouth corner stretch	50	40603
Protrusion	Upper lip retraction, lower lip retraction	30	21332
Jaw	Jaw thrust, jaw shift	30	21332

support the parishioners who so desire"). It can be observed that the trajectories produced by the Cohen-Massaro and Öhman models are smoother than the ANN-based models, since the former two models only produce one gesture per segment, while the ANN's have no such restrictions, although they are stabilised to some degree by the recurrent connections in the hidden layers. Trajectory jaggedness was not considered disturbing in the resulting animations.

The RMSE (root mean squared error) between target (**y**) and predicted (**z**) parameter trajectories were calculated over the test set for each parameter for all four models. The RMSE values were calculated as percentages of the full range of the target parameter. The analysis was carried out over three subsets of the test corpus: vowel segments, consonant segments and all segments except silence. It has been suggested that RMSE can be misguiding when comparing articulatory trajectories, since the RMSE is strongly influenced by areas of large amplitude (where larger errors are more likely to occur), whereas small deviations that could be crucial for e.g. lip closure could be overlooked. The Pearson product-moment correlation may give a better estimate of global match between signal shapes (Yehia et al., 1998). Table 2 shows the RMSE and the Pearson product-moment correlation (ρ_{yz}) for the four models, averaged over all parameters.

As the table indicates, the Cohen-Massaro model performs the best, having the lowest

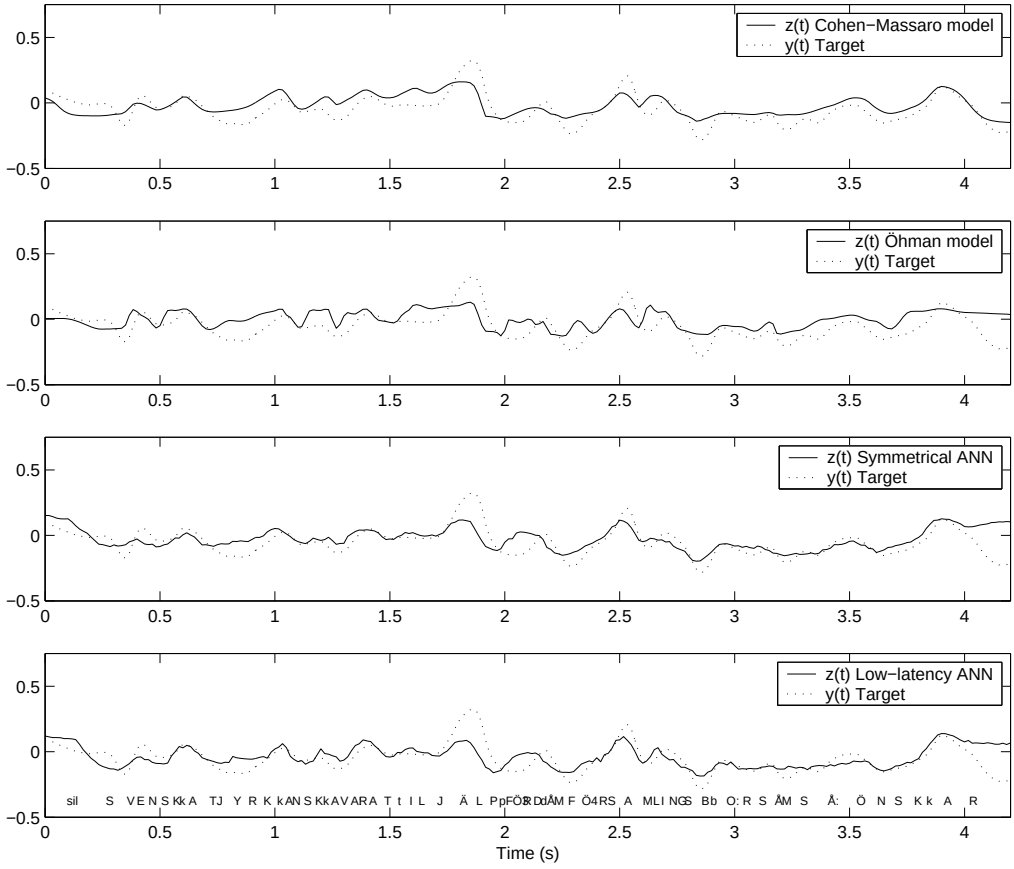

Figure 5. Target and estimated trajectories of the jaw rotation parameter for the four control models. Phonetic transcription in the Swedish *STA*-alphabet is given in the bottom plot.

Table 2. Average RMS error and correlation between target and estimated trajectories

	Control model							
	Cohen-Massaro (*3800 free parameters*)		Öhman (*1040 free parameters*)		ANN 1 (*123870 free parameters*)		ANN 2 low latency (*123870 free params*)	
	RMSE (%)	ϱ_{yz}	RMSE (%)	ϱ_{yz}	RMSE (%)	ϱ_{yz}	RMSE (%)	ϱ_{yz}
All segments	9,04	*0,6626*	9,50	*0,6188*	9,61	*0,6342*	9,61	*0,6065*
Vowels	9,06	*0,6631*	9,74	*0,6015*	9,86	*0,6323*	9,84	*0,6129*
Consonants	9,02	*0,6516*	9,34	*0,6156*	9,43	*0,6253*	9,45	*0,5926*

average RMSE as well as the highest correlation coefficient. Öhman's model shows slightly lower RMSE than the ANN models, although the correlation values are higher for the ANN 1 than Öhman's model. ANN 2 (low latency) shows almost identical RMSE-values as ANN 1 although the correlations are lower for the low latency model. In general, differences between the models are small, and it is difficult to say that one is better than another based on these numbers alone.

The table also lists the total number of free parameters, which differs significantly between the models. In general, a model with a large number of free parameters runs the risk of being over-trained to a particular data set, but by terminating the training before the error over the test set started increasing as a result of poor generalisation, this risk was controlled for. Furthermore, for the ANN's, many connection weights are small. The weights above 0,05 represent only 5% of the connections (about 6000 in total), a number that approaches the number of free parameters of the other models.

It can be noted that RMSE values agree reasonably well with those reported by Massaro et al. (in press), where the average RMSE over the training set (100 sentences) was trained was 13 % of the parameter range when 39 context independent phones were used, and 6% with 509 context-dependent phones.

6 PERCEPTUAL EVALUATION

While the objective measures are informative about how well the different control models predict the parameter trajectories, it is not obvious how they relate to the quality of the resulting animations. In order to obtain a rating of this, a perceptual evaluation was carried out.

6.1 Method

A sentence intelligibility test was carried out with 25 normal hearing native Swedish subjects.

Each of the four control models was used to synthesise animations with the animated talking head for a corpus of 90 phonetically labelled sentences not part of the training or test corpora, spoken by a male talker. The corpus consisted of short everyday sentences seven to nine syllables in length and has been developed specifically for audio-visual intelligibility testing by G. Öhngren, based on the work by MacLeod and Summerfield (1990).

In addition to the four data-driven control models, a rule-based control model (Beskow, 1995) and an audio-alone condition was included in the evaluation, yielding six presentation conditions. The frame rate of the animation was 30 Hz.

The acoustic signal was processed using a noise-excited vocoder (Shannon et al., 1995) with three frequency bands in the range of 100-5000 Hz. This form of audio degradation has been used in previous intelligibility studies (Siciliano et al., 2003) and has the advantage over additive noise of being robust to intensity perturbations in the speech signal.

Table 3. Average intelligibility scores for the 25 subjects for each condition.

	Audio-visual condition					
	Audio only	Cohen-Massaro	Öhman	ANN 1	ANN 2 (low latency)	Rule-based
Keywords correct (%)	62,7	74,8	75,3	72,1	72,8	81,1

Subjects were seated in front of a computer display and a loudspeaker, where they were presented with sentences and were asked to repeat what they perceived. Sentences were organised in six sub-lists of 15 sentences, each of which was paired with one of the six conditions. The pairing of lists and conditions was rotated between subjects, and the order in which the conditions occurred was randomised for each subject. Before the session, each subject was given a number of practice sentences for which the text of the sentence was given. In addition, each session began with 15 audio-alone sentences that were not scored.

6.2 Results

Results were scored by counting the percentage of correctly identified keywords for each 15-word list, where three words in each sentence had been defined as keywords. These percentages were entered into a repeated measures ANOVA, with the repeated variable being the six levels of the presentation condition. The effect of presentation condition was significant ($F(5, 120) = 9{,}13$; $p < 0.05$).

The average proportion of correct keywords for each condition is given in table 3. Pairwise comparisons using LSD (least significant difference) indicate that all face conditions give significantly higher intelligibility than the audio-alone condition, with $p <$ 0,05. Furthermore, the rule-based control model provides higher intelligibility than the data-driven ones, but no significant difference could be found between the four data-driven models on that significance level.

7 DISCUSSION

As stated in the introduction, the aim of this study was to find out whether there was reason to claim that any one of the data-driven control models is in some way superior to the others. The results of the objective as well as the perceptual evaluations indicate that this is not the case – all models perform more or less equally well, which means we are free to choose a model based on other criteria. As stated before, real-time considerations can be one such criterion, which makes the low-latency ANN model a strong candidate.

One question that demands an answer is why the data-driven models fall short of the rule-based model in the perceptual evaluation. This is not as surprising as one might first think. The rule-based model was developed with clear articulation and high intelligibility as the primary goal, and as such it almost tends to hyper-articulate. The data-driven models on the other hand, are trained to mimic the speaking style of the target speaker, who could be characterised as having a rather relaxed pronunciation. It should also be noted that the speaker was not selected on the basis of maximal visual intelligibility. So while the data-driven models are indeed capable of producing rather natural looking speech animations, they do not provide optimal intelligibility. It is likely that re-training the models on a corpus with a highly intelligible speaker would improve this aspect, but that is a matter of further investigation. While improving the intelligibility of a rule-based articulation scheme can be a very laborious process, a data-driven model can be automatically trained to more intelligible articulation, or to any particular style of speaking, which is one of the main attractions of employing data-driven techniques for generating synthetic visual speech.

8 ACKNOWLEDGEMENTS

The author would like to thank Bertil Lyberg for providing access to the Qualisys measurement facility. 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. Part of this work was supported by the IST-project SYNFACE.

9 REFERENCES

Aglefors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E & Öhman, T. (1998). Synthetic faces as a lipreading support, *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, pp. 3047-3050.

Allen, J., Hunnicut, M. S. and Klatt, D. (1987). *From text to speech: The MITalk system*. Cambridge, MA. Cambridge University Press.

Beskow, J. (1995). Rule-based Visual Speech Synthesis. *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*. Madrid, Spain, pp. 299-302.

Beskow, J. (1997). Animation of Talking Agents. *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'97)*, Rhodos, Greece, pp. 149-152

Bladon, R. A. & Al-Bamerini, A. (1976). Coarticulation resistance in English /l/. *Journal of Phonetics, 4*, 137-150.

Bregler, C, Covell, M. and Slaney, M. (1997) Video Rewrite: Driving Visual Speech with Audio. *Proceedings of ACM SIGGRAPH'97*, pp. 353-360.

Brooke, N. M. and Scott, D. S. (1998). Two- and Three-Dimensional Audio-Visual Speech Snthesis, *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp. 213-218.

Carlson, R., Granström, B. and Hunnicutt, S. (1982). A multi-language text-to-speech module. *Proceedings of the 7th International Conference on Acoustics, Speech, and Signal Processing (ICASSP'82)*, Paris, France, vol. 3, pp. 1604-1607.

Cohen, M. M. and Massaro, D. W. (1993). Modelling Coarticulation in Synthetic Visual Speech. Magnenat-Thalmann N., Thalmann D. (Eds), *Models and Techniques in Computer Animation*, Springer Verlag, Tokyo, pp. 139-156.

Cosi P., Magno Caldognetto E., Perin G. and Zmarich C. (2002) Labial Coarticulation Modeling for Realistic Facial Animation, *Proceedings of ICMI '02, 4th IEEE International Conference on Multimodal Interfaces 2002*, Pittsburgh, PA, USA, pp. 505-510.

Cole, R., Carmell, T., Connors, P., Macon, M., Wouters, J., de Villiers, J., Tarachow,, A., Massaro, D. W., Cohen, M. M., Beskow, J., Yang, J., Meier, U., Waibel, A., Stone, P., Fortier, G., Davis, A., & Soland, C. (1998). *Intelligent animated agents for interactive language training.* In STiLL: ESCA Workshop on Speech Technology in Language Learning, Stockholm, Sweden, pp. 163-166.

Ezzat, T, Geiger, G., Poggio, T. (2002). Trainable Videorealistic Speech Animation. *Proceedings of ACM SIGGRAPH 2002*, San Antonio, TX, pp. 388-398.

Gustafson, J. Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D. & Wirén, M. (2000). AdApt–a multimodal conversational dialogue system in an apartment domain. *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000).* Bejing, China, pp. 134-137.

Hällgren, Å. and Lyberg, B. (1998). Visual Speech Synthesis with Concatenative Speech. *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp. 181-183.

Le Goff, B., Guiard-Marigny, T., Cohen, M.M., Benoît, C. (1994). Real-time analysis-synthesis and intelligibility of talking faces. *Proceedings of the second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, New York, USA.

Le Goff, B. (1997). Automatic Modeling of Coarticulation in Text-to-Visual Speech Synthesis. *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97)*, Rhodos, Greece, pp. 1667–1670.

Löfqvist, A. (1990). Speech as audible gestures. In Hardcastle, W. J. and Marchal, A. (Eds.) *Speech Production and Speech Modelling.* Dordrecht: Kluwer Academic Publishers, pp. 289-322.

MacLeod, A. and Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24*, 29-43.

MacNeilage, P. D. & DeClerk, J. L. (1969). On the motor control of coarticulation in CVC monosyllables. *Journal of the Acoustical Society of America, 45*, 1217-1233.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, Massachusetts: MIT Press.

Massaro, D. W., Cohen, M.M., Tabain, M., Beskow, J. and Clark, R. (In press). Animated speech: Research progress and applications. In E. Vatiokis-Bateson, G. Bailly & P. Perrier (Eds.) *Audiovisual Speech Processing*, MIT Press.

Parke, F. I. (1982). Parametrized models for facial animation, *IEEE Computer Graphics, 2*(9), pp 61-68.

Pelachaud, C., Badler, N.I., and Steedman, M. (1996) Generating Facial Expressions for Speech, *Cognitive Science, 20* (1), 1-46.

Pelachaud, C. (2002). Visual Text-to-Speech. In Pandzic, I. & Forchheimer, R. (Eds.) *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, John Wiley & Sons, pp. 125-140.

Perkell, J. S. (1990). Testing theories of speech production: implications of some detailed analyses of variable articulatory data. In Hardcastle, W. J. and Marchal, A. (Eds.) *Speech Production and Speech Modelling.* Dordrecht: Kluwer Academic Publishers, pp. 263-288.

Press, W. H., Teukolsky, S. A., Vetterling, W., T. and Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing – 2nd ed.* Cambridge University Press.

Reveret, L., Bailly, G. and Badin, P. (2000). Mother: a New Generation of Talking Heads Providing a Flexible Articulatory Control for Video-Realistic Speech Animation. *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000).* Bejing, China, pp. 755-758.

Shannon, R. V., Zeng, F-G., Kamath, V., Wygonski, J. and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303-304.

Siciliano, C., Williams, G., Beskow, J. and Faulkner A. (2003). Evaluation of a Multilingual Synthetic Talking Face as a Communication Aid for the Hearing Impaired. To appear in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain.

Ström, N. (1997). Phoneme Probability Estimation with Dynamic Sparsely Connected Artificial Neural Networks. *The Free Speech Journal*, vol. 1(5).

Sumby, W. H., and Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, vol. 26, pp. 212-215.

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

Young S., Odell J., Ollason D., Valtchev V., and Woodland P. (1997). The HTK Book. Entropic Cambridge Research Laboratory.

Öhman, S. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America, 41*, 310-320.