# Training of Coarticulation Models using Dominance Functions and Visual Unit Selection Methods for Audio-Visual Speech Synthesis

*Zdeněk Krňoul, Miloš Železný, Luděk Müller, Jakub Kanis*

Department of Cybernetics
University of West Bohemia, Plzeň, Czech Republic
{zdkrnoul,zelezny,muller,jkanis}@kky.zcu.cz

## Abstract

This paper presents results of training of coarticulation models for Czech audio-visual speech synthesis. Two approaches for solution of coarticulation in audio-visual speech synthesis were used, coarticulation based on dominance functions and visual unit selection. For both approaches, coarticulation models were trained. Models for unit selection approach were trained by visualy clustered data. These data were obtained using decision tree algorithm. Outputs of audio-visual speech synthesis for both approaches were assessed and compared objectively.
**Index Terms**: audio-visual speech synthesis, talking head, coarticulation model

## 1. Introduction

In concatenation-based speech synthesis, basic speech units are concatenated to produce speech output. To obtain natural and intelligible speech, coarticulation effect has to be taken into account. Basic speech units are influenced by previously uttered speech units (backward coarticulation) and by speech units to be uttered after (forward coarticulation). If not properly handled, subsequent speech units do not fit to each other which produces unnatural effects at the basic speech unit boundaries.

This is true also (or even more) for audio-visual speech synthesis (parametrical talking head). Actually pronounced unit is visually presented as certain shape of lips and other articulatory organs. This shape can be obtained by analysis of real person articulation, parameterized and stored for future use in synthesis. The visual parameterization corresponds to the speech unit uttered, but is influenced by neighbouring units. First, it is affected by the shape of preceding unit due to inertia of articulatory organs. Thus, backward coarticulation has to take into account parameters of a preceding unit. Second, it is also affected by the unit that is to be pronounced consequently due to articulatory planning. Thus, forward articulation has to take into account parameters of next unit.

Coarticulation effect can be illustrated on difference in visual *t* in sequences *ata, oto, utu,* where *t* is so influenced by neighbouring vowels that it has almost no change in target from the vowel. On the opposite, *p* in sequences *apa, opo, upu* is also highly influenced by vowels, but due to lip closure has to reach the target in some parameters (closing lips). Compare these also with sequences *apu* or *upa* with different left and right context.

Coarticulation can be different for different languages. Our aim was to evaluate two methods for Czech audio-visual speech synthesis. Several approaches exist for solution of coarticulation in visual speech synthesis. From those, we selected two approaches. First is method of dominance functions proposed by Cohen and Massaro [1]. The idea is based on the theory of speech production [2]. This method uses target parameters for each visual speech unit and negative exponential functions to express influence on neighbouring visual speech units. Second method is the unit selection from clustered data. Similar method is used by Matoušek [3] in concatenation-based acoustic speech synthesis. Unit selection from phonetically clustered data was used also by Galanes et al. [4]. We use modification of this method using decision trees based on visual similarities.

## 2. Data preparation

The data driven models of coarticulation are necessary to train on data of particular speaker. The next condition is also the parameterization of visual speech. We use in our experiments the parameterization of lips (only outer contour) and jaw. This takes total 9 points in 3D (selection of 9 markers glued on the face, total dimension is 27). These geometric positions of points are obtained by noninvasive optical measurement. The observation of each of points produces one 3D trajectory with the rate 25 fps deinterlaced to 50 fps for visual part and 44 kHz for acoustic, both streams are time synchronized. All the trajectories approximate the movement and deformation of face surface. For designed purpose, we mainly localize the muscle *Orbicularis oris* and rotation of jaw.

The fact is that the amount of data from similar measurement is redundant. We have made the data reduction by principal component analysis (PCA). This method of reduction of dimension makes statistical analysis in direction of data deviance. We use singular value decomposition SVD. The hand-made observation of main component leads to dimension 3. We can find the shape of these components like: the first component (lip opening) by lower lisp and jaw, the second lip rounding with protrusion and the last upper lip rising. These parameters can be regarded as independent.

Czech language uses 42 phonemes and 5 non-speech events, total 47 units. In this designed experiment, we don't use relevant visemes subset because we want to model visemes as well as little divergences in each viseme subset. We obtained visual data for 3 speakers (1 female and 2 male) for Czech language. The female speaker is the professional speaker and 2 male speakers are students with nonprofessional articulation skills. The material is composed from 318 Czech sentences. The sentences were corrected by hand-made annotation and then automatically transcribed by phonetic transcription.

The text data was identical for all speakers and was collected from newspapers with resections good distribution of percent occurrence of phonemes and their context variants. The phonetic labeling of our data is made from synchronous

acoustic signal by well know viterbi-based forced alignment algorithm.

Recording of natural speech is a good way to analyze natural articulation. We recorded approximately one hour of natural speech per speaker. The advantage of our data is their direct usability for animation model. No extra transformations are needed. Animation model is created for every particular speaker using 3D reconstruction [5]. Geometric data in dimension 27 (9 points x 3 coordinates) correspond directly to the animation model [6] and in the best way control animation of a model and create visual speech.

## 3. Dominance functions

Cohen and Massaro [1] proposed a model based on Löfqvist [2] gestural theory of speech production. This coarticulation model describes the coarticulation by a dominance function. This dominance function describes the influence of the target parameter value for the actual speech unit on preceding and following visual speech units (visemes). Dominance function is based on a negative exponential function (1).

$$D = e^{-Ot^c} \quad (1)$$

This function falls from the center of the segment. This function is falling with running time $t$ from center segment. The rate of falling is driven by parameter $c$ and $O$. The parameter c modifies the steepness and $O$ modifies the rate of the falling. Cohen and Massaro [9] proposed function

$$D_{sp} = \alpha * e^{-Ot^c} \quad (2)$$

where $D_{sp}$ is the dominance of facial parameter $p$ on the speech segment $s$. The coefficient $\alpha$ drives the magnitude of the dominance function and thus drives the magnitude of influence on neighboring speech segments. We can see the complete dominance function in (2).

Cohen and Massaro used following alignment of the dominance function. Two dominance functions are defined for this parameter. Both functions start from centre of the speech segment. One function is falling along the running time and it models the forward coarticulation. Second function is falling against the running time and it models the backward coarticulation. The coefficient $\alpha$ is shared for both the functions and determines the amplitude of dominance function in the centre of the segment.

## 4. Visual unit selection

Unit selection method solves the coarticulation problem by slightly different way. Instead of storing only one representative speech unit for concatenative synthesis, more instances of each speech unit are stored. These units are clustered using decision trees and set of appropriate questions. The sequence of questions of the decision tree is trained using test set maximizing contribution of each decision of the tree.

These questions take into account usually left and rigtht context (preceding and following phoneme, affecting the actual one) and its properties (e.g. voiced/unvoiced, in visual case rounded/unrounded, etc.)

During synthesis, for each basic unit to be concatenated, the best fit unit is found using the trained decision trees and based on all information about the actual segment, its context, real durations for synthesized speech, etc. This approach is widely used in acoustic speech synthesis. In visual speech synthesis, unit selection is used rather rarely. One example is Galanes et al. [4]. However, they used the same set of questions for decision trees as for acoustic speech synthesis. But some phonetically similar segments can differ in visual expression and vice versa. We believe that unit selection approach would perform better if special "visual" set of questions is used.

## 5. Training of models

Training is based on recorded data. Using stored data we can determine part of trajectory for individual segments. Training is carried out for each PCA component separately. Question is how to best represent the speech segment (in our case a phoneme). In model based on dominance functions, every segment is represented by an articulation target. With help of further parameters, the continuous trajectory is generated from these targets (while the targets itself may not be reached).

We carried out data analysis and can coclude that every segment can be described by one value of articulation. It is not articulation target in its original meaning. It is one realization of given phoneme in context of continuous speech. The equivalent segment in the visual parameterization describes the shift of lip shape for these phonemes. The target positions is taken from the beginning of segment, in this place the lip shapes occur in full articulation "targets". This may be explained by the fact that the basic speech property is setting of articulatory organs before the acoustic realization of phoneme.This selection of target position appears to be usefull for the concatenate base synthesis of visual speech. In this way, the target value for each PCA component is prepared for each speech segment in the corpora.

We divided corpus data into training and testing part. Training part is created from first 270 sentences. In this part we have phonetically balanced percentual occurence of every phoneme including those very rare. Second, testing part is made from rest of sentences (48 sentences).

### 5.1. Training procedure – dominance functions

Model using dominance functions (Cohen-Massaro model) computes each point of a trajectory as a weighted sum of all articulation targets. The weights are given by the dominance (negatively exponential) functions. These functions fall with time from target peak (originally placed in center of segment) to both sides (forward and backward dominance). However, we use time demarcation obtained from acoustic data and thus we use beginning of segment instead of its center.

CM model has for each phoneme 5 unknown parameters (target value, dominance, forward and backward falling and the falling rate). We used only 4 parameters as in original model (not using the falling rate, considered as a global constant). Training of such number of free parameters is sped up by knowledge of a gradient of training equation. We computed these gradients according to Beskow [7]. Training is carried out by minimization of an error function that determines the quadratic error between synthesized and measured trajectory. The training process was verified on testing data.

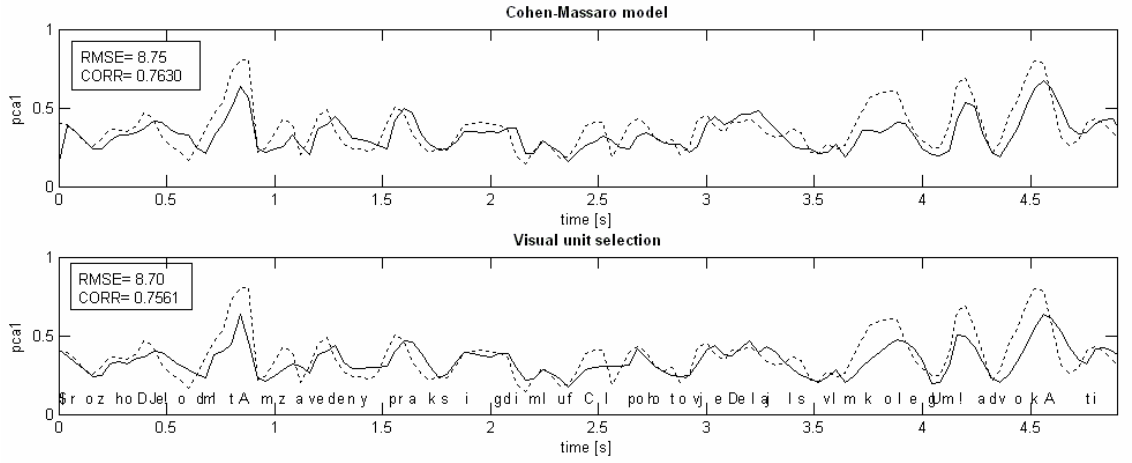


Figure 1 *Trajectories.*
*Real (dotted line) and synthesized (solid line) trajectory for female speaker and parameter "mouth opening" (PCA1) for czech sentence „Rozhodně odmítám zavedenou praxi, kdy mluvčí pohotově dělají svým kolegům advokáty."*

We trained 47 basic speech units for 4 articulation parameters for 3 trajectories, together 564 unknown values. Each PCA component trajectory could be trained independently – for 188 unknown values.

Process of training was stopped at the moment when error computed for training data stopped falling to avoid overtraining. Model was trained using truncated Newton bound constrained minimization with using for gradient information, which was implemented in the C language.

## 5.2. Training procedure – visual unit selection

The general idea is to take highest possible number of segment realizations for a particular phoneme and use them for synthesis. Our approach uses the same phoneme segments as previous model.

The aim of training of visual unit selection is to obtain the binary regression tree. The advantage is that the regression tree of particular phoneme should describe the most of variants of lip shape in the context other phonemes and have generalization property (aproximates even data (context) not seen during training). The three regression trees are construed from training data for each of 47 speech units.

Before training, we collected all appearance of targets for each phoneme, initial cluster. The number of phoneme appearances differs from tens to thousands. We choose for tree construction a phonetic and metrical context of segments and algorithm finds which condition can be the best for the split of the cluster. The splitting is made with regard to the minimization of mean visual distance (impurity). We use the Euclidean distance. The using of classification and regression tree (CART) techniques [8] provides suitable clustering of visual similar segments. Algorithm recursively applies the splitting until the minimal size of each cluster node (5 elements) is reached.

A regression tree walking is given by a specific sequence of questions that can answer binary value (yes or no) and returns the relevant set of response values. Each question asks if a requirement segment satisfies a given question. The requirement on segment context can be continuous or discrete. Depending on the answers to one question, you either proceed to another question or arrive at a fitted response value. The trees for some frequent phonemes are very large, that is why we use the tree pruning. Tree pruning is based on an optimal scheme that first prunes branches giving less improvement in error cost.

### *5.2.1. The questions*

Questions for construction of a decision tree are selected with an aim of maximum coverage of coarticulation effects that occur during concatenation of units at the phoneme level. The questions are specific for lip articulation. We selected only those variants that are meaningful for such articulation.

Questions are constituted in both discreet and continuous form. We selected such a set of questions that is most advantageous for this purpose. Discreet questions utilize the form known from phonetical decision trees, e.g. "What is left/right context of the segment?" Question can contain any subset of 47 phonemes and thus can be either more general ("Is the left context vowel?") or more specific ("Is the left context phoneme *a*?")

Next subset of questions is based on occurrence of closest coarticulatory resistent phoneme. The neighbourhood of the actual segment is searched for a dominant phoneme. From all 47 phonemes we selected specific subset of candidates that are coarticulation resistents (for example */pbm/, /fv/, vocals*).

Questions with continuous form of regression are constituted according to the time duration of neighbouring segments and also the actual segment itself. Such questions take into account the speed of a speech.

## 5.3. Synthesis

We take the labels of test data set for determining timing of synthesized trajectories. The phonetic context and time duration of synthesized segments are set to relevant regression tree. The found sequence of questions make tree walking from root to the target list value. The continuous trajectories are created by piecewise cubic interpolation method. This method takes the

Table 1. RMS errors and correlations for 3 speakers, 3 parameters and both compared models.

Dominance functions

| | RMSE [%] | | | | CORR | | | |
|---|---|---|---|---|---|---|---|---|
| speaker | PCA1 | PCA2 | PCA3 | average | PCA1 | PCA2 | PCA3 | average |
| 1 (male 1) | 8,93 | 7,90 | 7,74 | 8,19 | 0,8341 | 0,8453 | 0,7132 | 0,7975 |
| 2 (male 2) | 9,42 | 8,89 | 8,41 | 8,91 | 0,8028 | 0,8401 | 0,7202 | 0,7877 |
| 3 (female) | 8,75 | 7,10 | 8,96 | 8,27 | 0,7630 | 0,8444 | 0,6881 | 0,7652 |
| average | 9,03 | 7,96 | 8,37 | 8,46 | 0,8000 | 0,8433 | 0,7072 | 0,7835 |

Visual unit selection

| | RMSE [%] | | | | CORR | | | |
|---|---|---|---|---|---|---|---|---|
| speaker | PCA1 | PCA2 | PCA3 | average | PCA1 | PCA2 | PCA3 | average |
| 1 (male 1) | 10,99 | 9,10 | 8,62 | 9,57 | 0,7531 | 0,7990 | 0,6707 | 0,7409 |
| 2 (male 2) | 10,09 | 10,13 | 9,08 | 9,77 | 0,7691 | 0,7699 | 0,6409 | 0,7266 |
| 3 (female) | 8,70 | 7,65 | 8,73 | 8,36 | 0,7561 | 0,8186 | 0,7243 | 0,7663 |
| average | 9,93 | 8,96 | 8,81 | 9,23 | 0,7594 | 0,7958 | 0,6786 | 0,7446 |

best simulation of movements that are observed in the real data. The interpolation connects target position of neighboring speech segments and ensures the sampling rate of 50 fps.

## 6. Results

We carried out objective assessment of achieved results. We compared both coarticulatory models at the set of testing sentences. Results of the comparison can be expressed by various means. Graphical expression of difference between real and synthesized trajectory is depicted in Figure 1. Results for both coarticulation models are shown for female speaker and parameter PCA1 ("mouth opening").

To evaluate difference between the two coarticulation models we computed root mean square error (RMSE) between real and synthesized trajectory for each articulatory (PCA) parameter and for each speaker using corresponding testing data set. RMSE represents percentual deviation of given trajectory from given extent. RMSE reflects mostly errors that occur in higher amplitudes some small but important articulations may remain not noticed.

That is why we computed also Pearson's linear correlation coefficiens alike in [7] that can better assess total similarity (corrrectness) of the synthesized trajectory. Both RMSE and correlation figures are summarized in Table 1.

From these results it can be seen that visual unit selection (VUS) performs slightly worse than CM model (dominance functions), but still at about the same level. For speaker 3 and parameter PCA3 even VUS outperforms CM model from the point of view of correlation.

The problem of CM model is that the weighted sum of dominance functions does not ensure correct synthesis of targets, that are required to reach (such as the lip closure in bilabial stop for */pbm/* [7]). This property however is not readable in RMSE or correlation comparison. We proposed method that is free of this effect and has almost the same performance. Unlike in [4], we use questions based on visual features, which makes the decision trees more suitable for visual synthesis. Higher performance could be achieved by tuning the visually based decision tree questions.

## 7. Acknowledgements

This research was supported by the Grant Agency of the Academy of Sciences of the Czech Republic, project No. 1ET101470416.

## 8. References

[1] Cohen, M. M. and Massaro, D. W., "Modeling coarticulation in synthetic visual speech", In Thalmann, N. M. and Thalmann, D. (Eds.) *Models and Techniques in Computer Animation*, Tokyo : Springer-Verlag, 1993

[2] Löfqvist, A., "Speech as Audible Gestures", In, Hardcastle W. J. and Marchal A. (Eds.) *Speech, Production and Speech Modeling* Dodrecht : Kluwer Academic Publishers, 1990, pp. 289-322.

[3] Tihelka, D., Matoušek, J., "Unit selection approach for the Czech TTS system.", In *The 8th world multi-conference on systemics, cybernetics and informatics.* Orlando : International Institute of Informatics and Systemics, 2004. pp. 465-470.

[4] Galanes, F., M., Unverferth, J., Arslan, L., Talkin, D., "Generation of Lip-Synched Synthetic Faces from Phonetically Clustered Face Movement Data", In *International Conference on Auditory-Visual Speech processing,* Terrigal : AVSP, 1998

[5] Krňoul, Z., Železný, M., Císař, P., "Face model reconstruction for Czech audio-visual speech synthesis." In *SPECOM´2004*. Saint-Petersburg : SPIIRAS, 2004. pp. 47-51.

[6] Krňoul, Z., Železný, M., "Realistic face animation for a Czech Talking Head.", In *Text, speech and dialogue.* Berlin : Springer, 2004. pp. 603-610.

[7] Beskow, J., "Trainable Articulatory Control Models for Visual Speech Synthesis", Journal *of Speech Technology 7(4)*, pp. 335-349.

[8] Breiman L. et al. Classification and Regression Trees, Wadsworth Inc. Group, Belmont, Kalifornia, 1984