EG UK Theory and Practice of Computer Graphics (2010)
John Collomosse, Ian Grimstead (Editors)

Comparison of different types of visemes using a constraint-based coarticulation model

Oscar M. Martinez Lazalde Steve Maddock
University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello Street, Sheffield, S1 4DP, U.K.
+44(0) 114 2221800
{o.lazalde, s.maddock}@dcs.shef.ac.uk

ABSTRACT

A common approach to producing visual speech is to interpolate the parameters describing a sequence of mouth shapes, known as visemes, where visemes are the visual counterpart of phonemes. A single viseme typically represents a group of phonemes that are visually similar. Often these visemes are based on the static poses used in producing a phoneme. In this paper we investigate alternative representations for visemes, produced using motion-captured data, in conjunction with a constraint-based approach for visual speech production. We show that using visemes which incorporate more contextual information produces better results that using static pose visemes.

1. INTRODUCTION

There are different approaches to visual speech synthesis: interpolative synthesizers based on visemes [2, 3], concatenative synthesizers based on the concatenation of dynamic pieces [5, 12, 13], learning-based synthesizers based on learning a mapping between audio and visual features using machine learning techniques [4, 22] and physically-based synthesizers based on the physical simulation of the skin and muscles [1, 11]. Our approach is based on viseme interpolation, where the parameters describing a sequence of facial postures are interpolated to produce animation. For general facial animation, this approach gives artists close control over the final result, albeit at the expense of endless tweaking of facial posture to give the desired effect, and for visual speech it fits easily with the phoneme-based approach to producing speech.

An issue that visual speech synthesizers must deal with is coarticulation, which is the effect of context on a phoneme or its equivalent viseme. For example, in the word ‘boot’ the /o/ affects the lip shape of the /b/ and the /t/. For a concatenative synthesizer, coarticulation within pieces is accounted for, since the dynamic movement of the mouth is part of the representation, although there is still an issue at joins between pieces. For a viseme-based solution, the interpolation process must reproduce the effect. Here, the most commonly used approach is based on dominance functions [6]. Instead, we base our solution on a constraint-based approach (similar to [8, 9]), which formulates the coarticulation problem as an optimization problem. A viseme is treated as a cluster or range of poses centred on an ideal target pose. An objective function tries to fit a curve so that it passes through the viseme centre, but a set of constraints prevents this, so that the curve passes through the cluster, remaining in the viseme's range.

This paper investigates a number of different representations for a viseme that can be used in the constraint-based approach. These are formulated by considering the variability of a particular viseme, in isolation and in context, and considering the acceleration characteristics between a sequence of visemes. This kind of data is collected for a particular speaker by using a corpus (as described in Section 3), and the process of calculating the relevant parameters to be used in the animation process is known as tuning the model. For example, if the parameters for the viseme /t/ were known, this could be used in synthesizing the words 'dramatically' and 'dormitory'. However, the tuning process would need to take into account the context, and not assume that a /t/ behaves the same when between /a/ and /i/ as when between /i/ and /o/. Using a corpus to produce visemes makes the process similar to other data-driven approaches, such as concatenative approaches, although less data is stored for the viseme representation.

Section 2 will present an overview of the constraint-based approach. Section 3 describes the corpus recorded. Section 4 presents the type of data extracted from the corpus and how it is used to define the different types of visemes. Section 5 presents the results. Finally, section 6 presents conclusions and suggestions for future work.

2. CONSTRAINT-BASED VISUAL SPEECH

A posture (viseme) for a phoneme is variable within and between speakers. It is affected by context (the so-called coarticulation effect), as well as by such things as mood and tiredness. This variability needs to be encoded within the model. Thus, a viseme is regarded as a distribution around an ideal target. The aim is to hit the target, but the realisation is that most average speakers do not achieve this. Highly deformable visemes, such as an open mouthed /a/, are regarded as having larger distributions than closed-lip shapes, such as an /m/. Each distribution is regarded as a constraint which must be satisfied by any final speech trajectory. As long as the trajectory stays within the limits of each viseme, it is regarded as acceptable, and infinite variety within acceptable limits is possible.

To prevent the ideal targets from being met by the trajectory, other constraints must be present. For example, a global constraint can be used to limit the acceleration and deceleration of a trajectory. In practice, the global constraint and the

DOI: 10.2312/LocalChapterEvents/TPCG/TPCG10/199-206

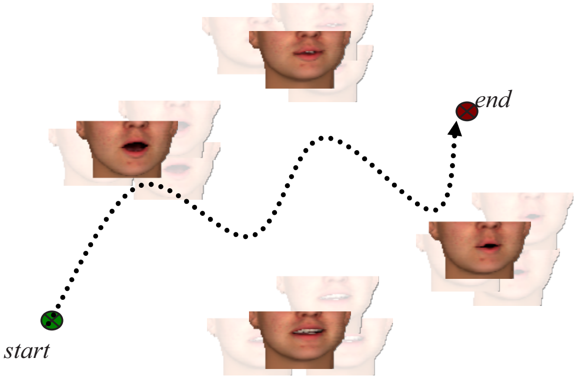

Figure 1. Conceptual view of the interpolation process through or near to clusters of acceptable mouth shapes for each viseme

distribution (or range) constraints produce an equilibrium, where they are both satisfied. Variations can be used to give different trajectories. For example, low values of the global constraint (together with relaxed range constraints) could be used to simulate under-articulation (e.g. mumbling). In addition, a weighting factor can be introduced to change the over/under articulation of a particular viseme.

Using the constraints, an optimisation function is used to create a trajectory that tries to pass close to the centre of each viseme. Figure 1 gives a conceptual view of this. We believe this approach better matches the mental and physical activity that produces the coarticulation effect, thus leading to better visual speech. In using a constrained optimisation approach [8], we need two parts: an objective function, Obj(X) and a set of bounded constraints C_j:

$$\begin{array}{ll} minimize & Obj(X) \\ subject\ to & \forall j: \underline{b_j} \le C_j(X) \le \overline{b_j} \end{array} \quad (2.1)$$

where $\underline{b_j}$ and $\overline{b_j}$ are the lower and upper bounds respectively. The objective function specifies the goodness of the system state *X* for each step in an iterative optimisation procedure. The constraints maintain the physicality of the motion.

The particular optimisation function we use is based on [8, 9]. However, whilst that work uses a system of weights to indicate the importance of particular targets, we add variability around the targets V_i using a constant, k_i, which represents over/under-articulation:

$$Obj(X) = \sum_i (S(t_i) - (V_i + a_i k_i))^2 \quad (2.2)$$

$$k_i = \begin{cases} \overline{V}_i - V_i & 0 \le a_i \le 1 \\ V_i - \underline{V_i} & -1 \le a_i < 0 \end{cases} \quad (2.3)$$

where $\underline{V_i}$ is the minimum value of V_i, $\overline{V}_i$ is the maximum value of V_i and a_i varies the amount of over-articulation and under-articulation. The objective function uses the square difference between the speech trajectory S and the sequence of ideal targets (visemes) V_i given at times t_i.

Table 1. Boundary Constraints

Constraints	Action
$S(t_{start}) = \varepsilon_{start}$	Ensures trajectory starts at ε_{start}
$S(t_{end}) = \varepsilon_{end}$	Ensures trajectory ends at ε_{end}
$S(t_{start})' = S(t_{end})' = 0$	Ensures the velocity is equal to zero at the beginning and end of the trajectory
$S(t_{start})'' = S(t_{end})'' = 0$	Ensures the acceleration is equal to zero at the beginning and end of the trajectory

The objective function is subject to a set of constraints. A speech trajectory S will start and end with particular constraints, e.g. a neutral state such as silence. These are the boundary constraints, as listed in Table 1, which ensure the articulators are in the rest state. If necessary, these constraints can also be used to join trajectories together. Range constraints are used to ensure that the trajectory stays within a certain distance of each target:

$$S(t_i) \in [\underline{V_i}, \overline{V}_i] \quad (2.4)$$

where $\underline{V_i}$ and $\overline{V}_i$ are, respectively, the lower and upper bounds of the ideal targets V_i. Finally, acceleration constraints are used to prevent the ideal targets being met, i.e. if Equation 2.4 and Table 1 are used in Equation 2.2, the ideal targets V_i will simply be met.

Two kinds of acceleration constraint are considered. Local acceleration constraints are used to limit the acceleration between two targets. A global acceleration constraint is used to dampen the trajectory. The local parametric acceleration is limited as follows:

$$|S(t)''| \le \lambda \qquad \text{where} \qquad t \in [t_i, t_{i+1}] \quad (2.5)$$

where λ is the maximum allowable magnitude of acceleration between two targets of the trajectory. If two consecutive targets are similar, the acceleration of the segment between them is small. If these targets are different, constraining the acceleration of the segment between them will make them similar. When a target has a coarticulation effect over its next target these two targets are similar and the acceleration of the segment between them is small.

The global parametric acceleration of a trajectory is limited as follows:

$$|S(t)''| \le \gamma \qquad \text{where} \qquad t \in [t_{start}, t_{end}] \quad (2.6)$$

where γ is the maximum allowable magnitude of acceleration across the entire trajectory. As this value tends to zero, the trajectory cannot meet its targets. As the global constraint is reduced the trajectory will eventually reach the limit of at least one range constraint.

A further constraint is added to ameliorate the effect of trajectory turning points occurring between viseme targets, since this can make the mouth movement appear to be out of synchronization with the audio. Local constraints that restrict the acceleration to zero at each target are used:

$$S(t_i)'' = 0 \qquad \forall t_i \in T \quad (2.7)$$

The speech trajectory S is represented by a cubic non-uniform B-spline. This gives the necessary C^2 continuity to enable Equations 2.5 and 2.6 to be applied. The optimisation problem is solved using a variant of the Sequential Quadratic Programming (SQP) method (see [24]). The SQP algorithm requires the objective function described in Equation 2.2. It also requires the derivatives of the objective and the constraints functions: the

Hessian of the objective function H_{obj} and the *Jacobian* of the constraints J_{cstr}. This algorithm follows an iterative process with the steps described in Equations 2.8, 2.9 and 2.10. The iterative process finishes when the constraints are met and there is no further reduction in the optimisation function (see section 5 for a discussion of this).

$$\Delta X_{obj} = -H_{obj}^{-1} \begin{pmatrix} \frac{\partial Obj}{\partial X_1} \\ \vdots \\ \frac{\partial Obj}{\partial X_n} \end{pmatrix} \qquad (2.8)$$

$$\Delta X_{cstr} = J_{cstr}^{+}(J_{cstr}\Delta X_{obj} - C) \qquad (2.9)$$

$$X_{j+1} = X_j + (\Delta X_{obj} + \Delta X_{cstr}) \qquad (2.10)$$

3. CAPTURING DATA

In order to produce specific values for the range constraints described in the previous section, we need to define the visemes that are to be used and measure their visual shapes on real speakers. In English, there is no formal agreement on the number of visemes to use. For example, Massaro defines 17 visemes [19] and both Dodd and Campbell [7] and Tekalp and Ostermman [23] use 14 visemes. We use 14 visemes for Mexican-Spanish [18], as listed in table 2. Many of these are similar to the English visemes, although there are exceptions. The phoneme /v/ is an example where there is a different mapping between Spanish and English visemes; in English speech the phoneme maps to the /F/ viseme whereas in Spanish, the /v/ phoneme corresponds to the /B/ viseme. There are also letters, like /h/, that do not have a corresponding phoneme in Spanish (they are not pronounced during speech) and thus have no associated viseme. Similarly, there are phonemes in Spanish that do not occur in English, such as /ñ/, although there is an appropriate viseme mapping, in this example to the /N/ viseme.

The data for the visemes in Table 2 was captured both statically and dynamically. In the static case, each speaker performed (only once) the static mouth shapes of each of the representative phonemes in Table 2. In the dynamic case, two native Mexican-Spanish speakers (one female and one male) were recorded using an inexpensive mocap system built using two high speed (Casio Exilim F1) cameras. Both speakers recorded a corpus of VCV combinations as described in Table 3 where C is any of the representative phonemes in Table 2 and V are the viseme vowels a, e, i and o. Each of the VCV segments was repeated 10 times. The speed of production of the segments was controlled by using a metronome. The speed was measured by syllables per minute (spm) and the recorded speeds used in this work were 100 spm and 200 spm. This gives a total of 2880 segments per speaker (16 VCV segments multiplied by 10 repetitions multiplied by 9 consonants multiplied by 2 speeds). Syllables per minute were chosen as it is easy to synchronize a syllable to a beat than a single phoneme as some consonants are hard to keep for long periods of time, e.g. plosives. When synchronizing to the metronome, the first vowel would align with a metronome beat, then the consonant and the second vowel align with the following beat. This means the first vowel is longer than the second. To make the vowels even in length, a /t/ was introduced at the beginning and end of the segment. The /t/ was selected as it is a consonant that is easily influenced by surrounding vowels, e.g. consider the words steed and boot.

Table 2. Mexican-Spanish viseme definition

Phonemes: IPA and example word	**Representa tive phoneme**	**Viseme name (vowels lower case, consonants upper case)**
silence		NEUTRAL
/a/c**a**sa	/a/	a
/b/**b**eber, /m/**m**arcar, /p/**p**artir	/b/	B
/tʃ/**ch**orro, /ʎ/**ll**uvia	/tʃ/	CH
/d/**d**edo, /s/**s**ol, /t/**t**odo	/d/	D
/e/p**e**so	/e/	e
/f/**f**also	/f/	F
/i/s**i**	/i/	i
/x/relo**j**,/g/**g**anar	/x/	J
/k/**c**asa	/k/	K
/n/**n**iño, /ɲ/ni**ñ**o	/n/	N
/o/b**o**sque, /u/c**u**na	/o/	o
/ɾ/pe**r**o, /r/pe**rr**o	/r/	R
/l/**l**os	/l/	L

Table 3. Vowel-Consonant-Vowel (VCV) combinations

Vowels	**a**	**e**	**i**	**o**
a	aCa	aCe	aCi	aCo
e	eCa	eCe	eCi	eCo
i	iCa	iCe	iCi	iCo
o	oCa	oCe	oCi	oCo

The recorded mocap data is in the form of the 3D coordinates of the markers placed on the speaker's face. These markers are used to deform a 3D synthetic face model according to the speaker movement so the 3D model replicates the movement . We call this mapped 3D data. This is done for each frame in a recorded VCV segment using Radial Basis Functions [21] and Mixtures of Probabilistic Principal Component Analysis [10]. We use Principal Components (PCs) from Principal Component Analysis (PCA) as the parameterization [14, 17]. The PCs are obtained by applying PCA to a set of 17 different poses obtained from FaceGen (Singular Inversions: http://www.facegen.com/) and the first seven PCs are used as the parameterization. By projecting the mapped 3D data to the space defined by the PCs a set of seven PC parameter curves is obtained. A frame were each of the visemes is thought to be produced is extracted. The variability of a viseme is obtained by grouping the frames were the given viseme was produced. This data is then used with an interpolation process to create a synthetic animation curve to compare against the PC parameter curves from the 3D mapped data.

4. EXTRACTING DATA AND VISEME TYPES

In order to extract values for the visemes the recorded segments have to be labelled. Labelling identifies the instant in time the viseme is executed. Once labelled, all the executions of a given viseme can be used to give a range for it. Different types of data

Table 4. The experiments

Experiment number	Viseme type	Range type (max, min)	Interpolation Method
0	Static	None	spline
1	Static	Type 1	constraints
2	Static	Type 2	constraints
3	Static	Type 3	constraints
4	Coarticulated Type 1	None	spline
5	Coarticulated Type 2	None	spline
6	Coarticulated Type 3	None	spline
7	Coarticulated Type 1	Type 1	constraints
8	Coarticulated Type 2	Type 2	constraints
9	Coarticulated Type 3	Type 3	constraints
10	Enhanced: Static	Type 1	constraints
11	Enhanced: Static	Type 2	constraints
12	Enhanced: Static	Type 3	constraints
13	Enhanced: Coarticulated Type 1	Type 1	constraints
14	Enhanced: Coarticulated Type 2	Type 2	constraints
15	Enhanced: Coarticulated Type 3	Type 3	constraints

can be extracted from the corpus, which we define as Type 1, 2, 3 and 4. In addition, Type 0 data is the visemes captured while the speaker is holding the production of a viseme. These are essentially a baseline since they contain no context and thus no coarticulation information. We will refer to these as static visemes.

In Type 1 data, a coarticulated viseme is taken as the mean of all the repetitions of the viseme. The viseme range is also taken from all the repetitions. For the vowels, this means mixing the repetitions of a given vowel when it appears at the beginning of a VCV segment and when it appears at the end of a VCV segment for all consonants C. For the consonants, the consonant data is obtained by mixing all the consonants of all the combinations of VCV where the consonant is used.

In Type 2 data, the coarticulated visemes and the viseme ranges for consonants are obtained in the same way as in Type 1 data. For vowels, they are obtained by separating the order in which the vowel appears. The vowel will have a coarticulated viseme and a viseme range for when it starts a VCV segment and another for when it ends the segment, for all consonants.

In Type 3 data, the coarticulated viseme and its range for a vowel depend on the position of the vowel, on the consonant used and also on the vowel used before or after the consonant. For example, the vowel /a/ preceding the consonant /m/ will have different values for the coarticulated viseme value and for the viseme range when appearing in ‘ama’ than in ‘amo’. The vowel /a/ will have a different coarticulated viseme value and viseme range when appearing after the consonant /m/ when appearing in ‘ema’ than when appearing in ‘ima’. For the consonants, the coarticulated viseme and the viseme range will be defined according to the vowel that is before and after the consonant. For example, the consonant /m/ will have a different coarticulated viseme and viseme range when appearing in ‘ama’ than when appearing in ‘ame’.

Type 4 data are acceleration ranges. These acceleration ranges depend on what is before or after a viseme. For example, the acceleration range between a vowel /a/ and a consonant /m/ will be different when appearing in ‘ama’ than when appearing in ‘ame’. Also the acceleration range between a consonant /m/ and a vowel /e/ will be different when appearing in ‘ame’ than when appearing in ‘ime’.

From this data, static visemes, coarticulated visemes and enhanced visemes are obtained. Static visemes are visemes captured while the speaker is holding the production of a viseme, and thus do not incorporate context. Coarticulated visemes are taken as the mean of a viseme produced several times during recorded speech. Three types of coarticulated visemes were extracted. They correspond to the Type 1, 2 and 3 data, respectively. Enhanced visemes are the combination of static visemes or coarticulated visemes with local acceleration information before and after the viseme. This information is obtained from Type 4 data.

5. EVALUATION AND RESULTS

The different kinds of viseme data are used in our constraint-based coarticulation model in order to produce synthetic animation curves. In addition, a simple spline curve fitted through the targets is used as a baseline for the experiments. Table 4 describes all the possible experiments. As an example, experiment 13 uses coarticulated Type 1 visemes to produce synthetic animation curves using the constraint-based approach.

An objective evaluation is then conducted by comparing the synthetic curves against the original recorded curves (from the mapped 3D data) using Root Mean Square (RMS). RMS is used as the synthetic and recorded curves are aligned. If the curves were not aligned, Dynamic Time Warping (DTW) [20] could be used. The first evaluation is done using the corpus used to extract the data, i.e. the synthetic curves are compared against the recorded data in the corpus. The second evaluation uses a set of words that are different to the data in the corpus.

For the first evaluation, each of the experiment settings was evaluated against each of the recorded VCV segments described in section 3. This means each of the experiment settings had to synthesize each of the different sentences and their repetitions. This evaluation was done just for the first PC animation parameter; a similar evaluation could be carried out for the rest of the parameters. Each experiment setting synthesized 1440 different VCV sentences (9 consonant visemes multiplied by 16 VCV combinations and 10 executions of each VCV performed by a given speaker at a given speed).

As mentioned in section 4, each of the executions of the VCV combinations was labelled. In each execution the timing is slightly different and the magnitude of each curve and any common points is slightly different. Given the labelling, 10 synthesized curves are produced for each VCV combination. The synthesized curves are compared with the respective original.

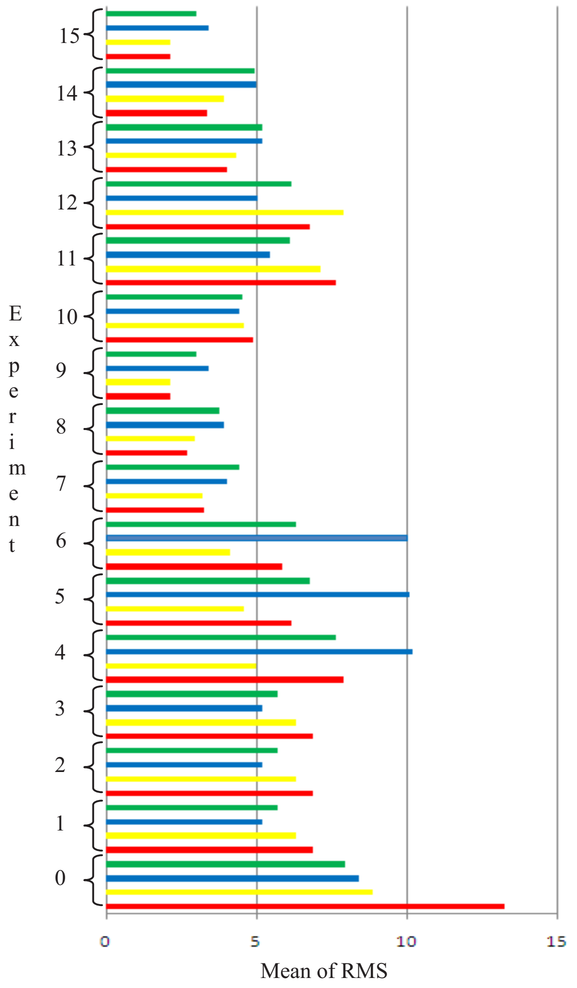

Figure 3. VCV evaluation results: Female speaker at 100 spm (red); Female speaker at 200 spm (yellow); Male speaker at 100 spm (blue); Male speaker at 200 spm (green).

Figure 3 shows the RMS mean for the evaluation of each experiment for each speaker. It can be observed that when using static visemes, the coarticulation model (experiments 1, 2, 3) performs better than the spline interpolation (experiment 0). This is because the coarticulation model constrains the trajectory to pass between the viseme ranges. Similar behaviour can be observed using coarticulated visemes.

Also, it can be observed that, in general, both interpolation approaches perform better using coarticulated visemes than static visemes. For example, for splines, experiments 7, 8 and 9 give better results than experiment 0. For the constraint-based model, experiments 13, 14 and 15 perform better than experiments 4, 5 and 6. This is clearly because coarticulated visemes are captured in continuous speech, i.e. in context, whilst static visemes are captured when holding a position and in isolation, and thus do not include coarticulation effects.

In general, it can be observed that enhanced visemes perform better than static visemes. Enhanced visemes contain coarticulation information as local acceleration constraints (obtained from recorded data) applied before and after the viseme. It also can be observed that coarticulated visemes perform better than static enhanced visemes.

Table 5. Results of ANOVA tests for VCV evaluation results.

Speaker	ANOVA	F	Significance
Female (100 spm)	F(21,33098)	1719.69	0.0001
Female (200 spm)	F(21,33098)	504.461	0.0001
Male (100 spm)	F(21,33098)	1125.64	0.0001
Male (200 spm)	F(21,33098)	279.279	0.0001

The best performance is reached by the coarticulated viseme with Type 3 data and the enhanced viseme with Type 3 data. This tells us that adding local acceleration constraints (to form the enhanced viseme) to coarticulated visemes of Type 3 data is redundant. It can also be observed that, in general, using Type 3 data produces better results than using Type 2 data, which, in turn, produces better results than using Type 1 data. This is because Type 3 data is more refined than Type 2 data, and Type 2 data is more refined than Type 1 data.

An Analysis of Variances (ANOVA) can give us confidence in what is observed in Figure 3. In order to apply ANOVA, a test on the homogeneity of variances is required [15]. Levene's test [16] was applied to each speaker's results at each capture speed and did not find any evidence for a departure from the homogeneity assumption (p=0.0001). A one-way ANOVA was applied on type of experiment (see Table 5) for each speaker at each speed. The ANOVA tests that the means are the same across the types of experiments. Table 5 shows that this has a small probability for any speaker's results at any speed with 95% of confidence. So, the differences between experiments are backed up by this test and these differences are not a random coincidence.

The experiment settings in Table 4 were used to synthesize longer words. The words were nonsense words in Mexican-Spanish constructed randomly and alternating vowels and consonants. This means a vowel is always preceded and followed by a consonant and a consonant is always preceded and followed by a vowel. The words used for this evaluation are shown in Table 6. The consonant visemes appear in upper case.

These words were performed by the male and female speaker at 100 spm and 200 spm and at a free normal speaking rate and recorded. The pronunciation of these words is the same for both speakers as both of them were pronounced by Mexicans and there is no ambiguity in pronunciation in Mexican-Spanish. As the corpuses consist of Vowel-Consonant-Vowel (VCV) sentences, information about ranges and viseme centres for consonants appearing between vowels (VCV) is available. Information about a given vowel between consonants is not available (CVC). The nonsense words in Table 6 contain vowel between consonants segments. For experiments using Type data 1 the vowel ranges and centres with no change were used. The following was done for experiments using Type 2 data:

- When the vowel is between consonants, the average of the range of the vowel after the first consonant and the range of the vowel before the second consonant is taken as the range. The same is done for the centre.

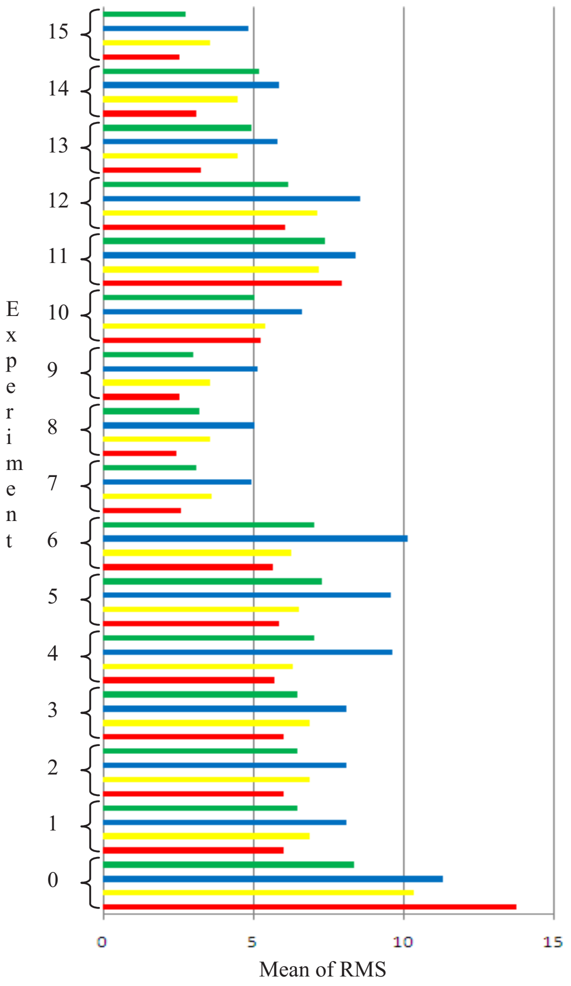

Figure 4. Word evaluation results: Female speaker at 100 spm (red); Female speaker at 200 spm (yellow); Male speaker at 100 spm (blue); Male speaker at 200 spm (green).

- When the vowel starts the word, the range of the vowel before the consonant is taken. The same is done for the centre.
- When the vowel ends the word, the range of the vowel after the consonant is taken. The same is done for the centre.

For experiments using Type 3 data the following was done:

- When the vowel is between consonants, the first consonant is not the first letter of the word and the second consonant is not the last letter of the word, the average of the range of the vowel given the vowel before the first consonant and the range of the vowel given the vowel after the second consonant is taken. The same is done for the centre.
- When the vowel is between consonants and the first consonant is the first letter of the word, an average is taken between all the ranges of the vowel given any vowel before the first consonant. All this is averaged with the range of the vowel given the vowel after the second consonant. The same is done for the centre.

Table 6. Words used for evaluation (vowels in lowercase, consonants in uppercase)

Word number	Word
1	JiBeKiRoNa
2	DiFaLeN
3	oCHiRo
4	CHeLa
5	aKeBoF

Table 7. Results of ANOVA tests for VCV evaluation results.

Speaker	ANOVA	F	Significance
Female (100 spm)	F(21,93)	5.602	0.0001
Female (200 spm)	F(21,93)	3.324	0.0001
Male (100 spm)	F(21,93)	1.471	0.108
Male (200 spm)	F(21,93)	2.998	0.0001

- When the vowel is between consonants and the second consonant is the last letter of the word, an average is taken between all the ranges of the vowel given any vowel after the second consonant. All this is averaged with the range of the vowel given the vowel before the first consonant. The same is done for the centre.
- When the vowel starts the word, the range of the vowel given the vowel after the consonant is taken. The same is done for the centre.
- When the vowel ends the word, the range of the vowel given the vowel before the consonant is taken. The same is done for the centre.

Local acceleration ranges were calculated in a similar way. Each of the experiment settings in Table 4 was used to synthesize the words in Table 6 for the male and female speaker at 100 and 200 spm. Figure 4 shows the RMS mean for the evaluation of each experiment for each speaker. As for the VCV evaluation, using coarticulated visemes or enhanced visemes gives better results than using static visemes. Using coarticulated visemes gives better results than enhanced visemes. Using coarticulated visemes with Type 3 data gives the same results as enhanced visemes with Type 3 data. Using the coarticulation model gives better results than using spline interpolation. Again, Levene's test and the ANOVA test were applied (see Table 7) to support these observations. In this case, the ANOVA test for the results for the male speaker at 200 spm is non-significant. This could be due to the need for more comparisons but these results can be taken as informal evidence.

To confirm visually the results of the objective comparison a few frames for the first PC parameter are shown in Figure 5. Figure 5 shows the visual results for the PC 1 parameter for experiments 0, 9 and 13 in comparison to the mapped 3D data for one of the executions of the VCV sentence 'taNat' for the male speaker at 200 spm. It can be observed that the experiment 9 is the closest visually to the mapped 3D segment. This matches the objective evaluation results.

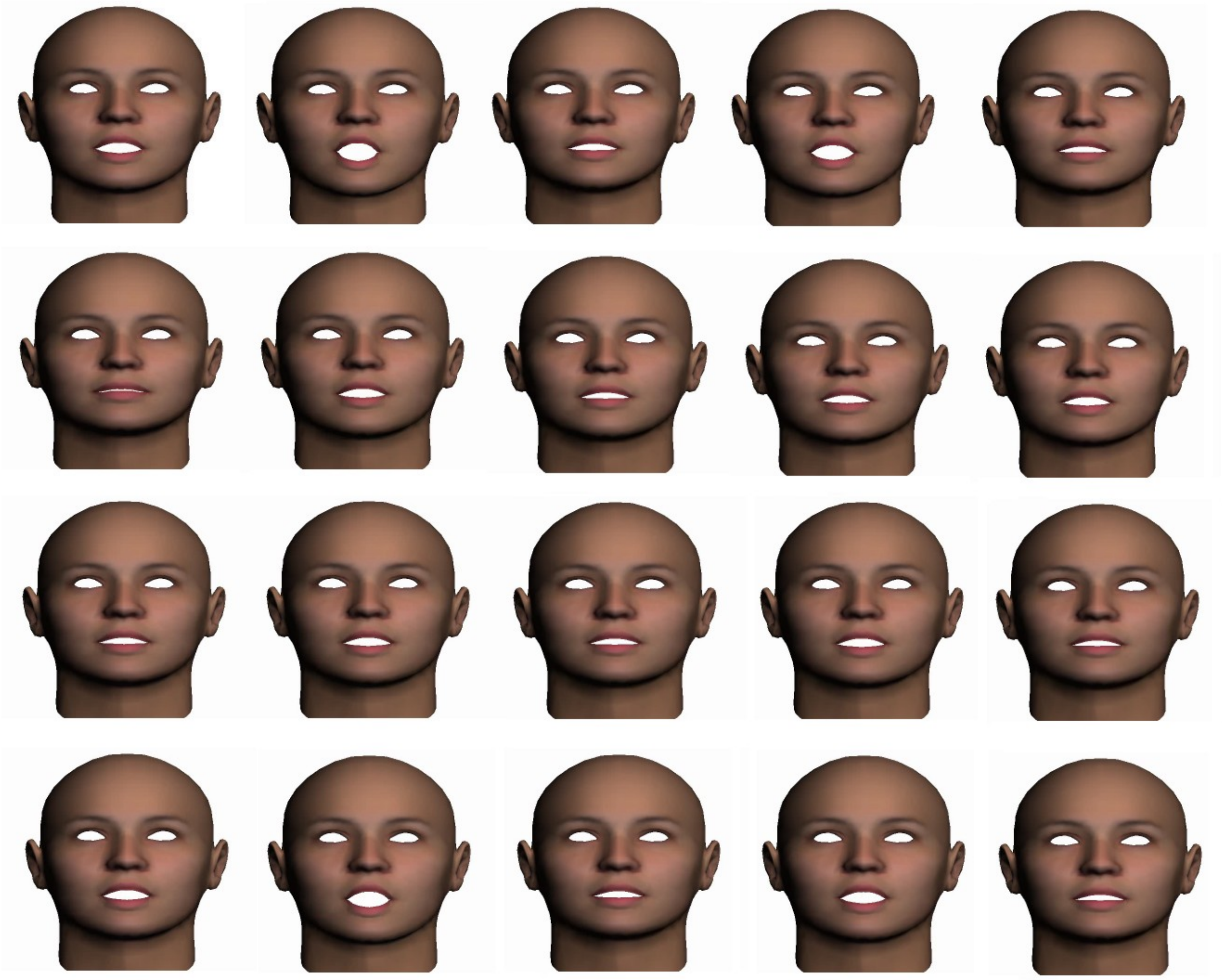

Figure 5. Visual comparison for PC 1 of mapped 3D data (first row) against experiments 0 (second row), 12 (third row) and 9 (fourth row) for the VCV segment 'taNat'. The first column corresponds to the first 't', the second column corresponds to the first 'a', the third column corresponds to 'N', the fourth column corresponds to the second 'a' and the last column corresponds to the second 't'.

6. CONCLUSIONS

We have presented a constraint-based coarticulation model for interpolative visual speech. Parameters of the constraint-based approach such as viseme centres, variability of visemes and acceleration between two visemes were tuned from data from two real speakers. An evaluation of the use of three different types of visemes (static, coarticulated and enhanced) for synthesising visual speech synthesis has been conducted. Using the constraint-based approach it was found that coarticulated visemes using the most refined data (Type 3 data: centre of a viseme and its variability depending on what viseme is before and what viseme is after) provide the best approximation to real recorded curves. Adding acceleration information to the coarticulated visemes of Type 3 data, resulting in enhanced visemes, gave similar results. Thus, the information added by the acceleration information offers no improvement to these coarticulated visemes.

When using these refined coarticulated visemes, a higher level of planning is necessary as the centre and range of the viseme are set according to what is before and after the viseme. We can argue that a higher level of planning is necessary for any coarticulation model in order to improve the results. This has repercussions on how coarticulation models are usually tuned. For example, for the Dominance Functions coarticulation model [6], the parameters of the Dominance function for a given viseme are usually calculated in two ways: from the interaction of the viseme in only one context (case A) and from the interaction of the viseme in several contexts (case B). In case A, for example, let's suppose the parameters for the viseme T (representing the visual counterpart of the phoneme /t/) are being tuned and this is done from the recorded word 'dramatically'.

This tuning would not produce a good trajectory when synthesizing the word ‘dormitory’. With the data we collected we found the centre of a viseme varies depending on what is before and after it. In case B, let’s suppose the tuning is done from several words by an optimization process: ‘dramatically’, ‘dormitory’, ‘petal’, ‘outer’. The tuning done for Dominance Functions is a more general tuning, and it is similar to the less refined data (Type 1 data) used with the constraint-based approach. However, we found that the most refined data (Type 3 data) gave better results. We can make the obvious summary that when more information is added, the results are better.

The amount of manual work required to extract the different types of data didn’t vary across type of data, so collecting the most refined data (which includes more information) implied the same work as collecting the less refined data. But, more work is needed for extracting this data than for collecting static visemes. We have yet to compare the effort in comparison to collecting data for a concatenative synthesizer. We believe it is less effort and that the use of a refined viseme in conjunction with a constraint-based approach potentially offers more control and flexibility in the production of visual speech.

Initial experiments with longer utterances (e.g. the concatenation of all the words in Table 6) have been carried out, and, again, coarticulated visemes using the most refined data (Type 3 data) provided the best approximation to real recorded curves. However, further evaluation of longer utterances needs to be done.

The evaluation presented in this paper concentrates on one animation parameter (PC1) which mainly describes gross mouth open-close movement. The other PCs contain more subtle mouth movements and further evaluation using those PCs needs to be done. Also, the synthesized 3D control points and ground truth ones could be compared rather than the PCs. RMS was used in the evaluation but other measures such as correlation coefficients could be investigated. Finally, evaluation using data from one speaker to synthesize sentences of the other speaker would provide further interesting work.

7. ACKNOWLEDGMENTS

Thanks to Monica Acosta who helped us in the recording process. Thanks to CONACYT for funding this work.

8. REFERENCES

[1] Albrecht, I. and Haber, J., 2002. Speech Synchronization for Physics based Facial Animation. *In: WSCG'02, 2002*. 9-16.

[2] Benoit, C. and Le Goff, B., 1998. Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. *Speech Communication,* **26** (1-2): 117-129.

[3] Bevacqua, E. and Pelachaud, C., 2004. Expressive audio-visual speech. *Computer Animation and Virtual Worlds,* **15** (3-4): 297-304.

[4] Brand, M., 1999. Voice Puppetry. *In: ACM SIGGRAPH'99, 1999*. 21-28.

[5] Cao, Y., Tien, W. C., et al., 2005. Expressive speech-driven facial animation. *ACM Transactions on Graphics,* **24** (4): 1283-1302.

[6] Cohen, M. and Massaro, D., 1993. Modeling coarticulation in synthetic visual speech. *In: Models and Techniques in Computer Animation, 1993 Tokyo, Japan*. 139-156.

[7] Dodd, B. and Campbell, R., 1987. Hearing by Eye: The Psichology of Lipreading. *British Journal of Ophthalmology,* **72** (6): 479.

[8] Edge, J. D., 2004. *Techniques for the Synthesis of Visual Speech*. Thesis (Phd). The University of Sheffield.

[9] Edge, J. D. and Maddock, S., 2003. Spacetime Constrains for Viseme-based Synthesis of Visual Speech. Technical Report, Technical Report CS-04-03 Department of Computer Science, University of Sheffield.

[10] Gunnarsson, O. and Maddock, S., 2008. Sketching Faces. *In: Fifth Eurographics Workshop on Sketch-Based Interfaces and Modelling 2008, 2008 Annecy, France*.

[11] Haber, J., 2003. MEDUSA - A Facial Modeling and Animation System. *Forschung und wissenschaftliches Rechnen - Beiträge zum Heinz-Billing-Preis,* (58): 13-28.

[12] Huang, F., Cosatto, E., et al., 2001. Triphone based unit selection for concatenative visual speech synthesis. *In: International conference on acoustics, speech, and signal processing Vol 2: Signal processing theory and methods, 2001 Salt Lake City, UT, USA*.

[13] Kshirsagar, S. and Magnenat-Thalmann, N., 2003. Visyllable Based Speech Animation. *Computer Graphics Forum,* **22** (3): 631.

[14] Kshirsagar, S., Molet, T., et al., 2001. Principal Components of Expressive Speech Animation. *In: International Conference on Computer Graphics, 2001*. IEEE Computer Society, 38-44.

[15] Landau, S. and Everitt, B., 2004. *A Handbook of Statistical Analyses using SPSS,* Chapman & Hall. 354 p.

[16] Levene, H., 1960. Robust test for the equality of variance. In *Contributions to Probability and Statistics*(Ed, Aikin, O.) Stanford University Press, Stanford, CA.

[17] Martinez-Lazalde, O., Maddock, S., et al., 2007. A Mexican-Spanish talking head. *In: CyberGames 2007, 2007 Manchester, UK*. 17-24.

[18] Martinez-Lazalde, O., Maddock, S., et al., 2008. A Constraint-Based approach to Visual Speech for a Mexican-Spanish Talking Head. *International Journal of Computer Games Technology,* **Volume 2008**: 7 pages.

[19] Massaro, D. W., 1998. *Perceiving talking faces: from speech perception to a behavioral principle,* MIT Press, Cambridge, Mass. p.

[20] Sakoe, H. and Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing,* **26** (1): 43-49.

[21] Sanchez, M., Edge, J. D., et al., 2003. Use and Re-use of Facial Motion Capture Data. *In: Vision, Video and Graphics 2003, 2003 University of Bath, UK*. 135-142.

[22] Tamura, M., Masuko, T., et al., 1998. Visual speech synthesis based on parameter generation from HMM: speech-driven and text-and-speech-driven approaches. *Proceedings of AVSP,* **1998**: 219-224.

[23] Tekalp, M. and Ostermann, J., 2000. Face and 2-D mesh animation in MPEG-4. *Signal Processing: Image Communications,* **15** (4): 387-421.

[24] Witkin, A. and Kass, M., 1998. Spacetime constraints. *In: 15th annual conference on computer graphics and interactive techniques, 1998*. 159-168.