

Automatic Viseme Clustering for Audiovisual Speech Synthesis

Wesley Mattheyses, Lukas Latacz and Werner Verhelst

Vrije Universiteit Brussel, Dept. ETRO-DSSP,
Interdisciplinary Institute for Broadband Technology IBBT, Brussels, Belgium

{wmatthey, llatacz, wverhels}@etro.vub.ac.be

Abstract

A common approach in visual speech synthesis is the use of visemes as atomic units of speech. In this paper, phoneme-based and viseme-based audiovisual speech synthesis techniques are compared in order to explore the balancing between data availability and an improved audiovisual coherence for synthesis optimization. A technique for automatic viseme clustering is described and it is compared to the standardized viseme set described in MPEG-4. Both objective and subjective testing indicated that a phoneme-based approach leads to better synthesis results. In addition, the test results improve when more different visemes are defined. This raises some questions on the widely applied viseme-based approach. It appears that a many-to-one phoneme-to-viseme mapping is not capable of describing all subtle details of the visual speech information. In addition, with viseme-based synthesis the perceived synthesis quality is affected by the loss of audiovisual coherence in the synthetic speech.

Index Terms: audiovisual speech synthesis, visemes, facial animation

1. Introduction

In the fields of visual speech analysis and synthesis, a common approach is to treat the speech as a sequence of so-called visemes. Visemes can be considered as the particular facial and oral positions that show when a speaker utters phonemes. Earlier research has indicated a many-to-one relation between phonemes and visemes [1] and many different phoneme-to-viseme mappings have been proposed. Most mappings are based on the articulatory and phonetic properties of the phonemes, in combination with subjective perception experiments to measure the visual confusability between different phonemes [2]. Despite the fact that the actual mapping between phonemes and visemes tends towards a many-to-many mapping due to visual co-articulation effects [3], many reports can be found where a many-to-one phoneme-to-viseme mapping has been successfully applied for automatic lip-reading as well as for visual speech synthesis [4].

In previous research [5] we designed an audiovisual text-to-speech synthesis (AVTTS) system that can construct a synthetic audiovisual speech signal based on an input text. We used a data-driven unit-selection synthesis approach where the system's database consists of original audiovisual speech recordings whose visual speech mode is modeled by active appearance models (AAMs)[6]. This allows to carefully analyze, select and concatenate original audiovisual speech segments. We have shown that in audiovisual speech perception, the synchrony and the coherence between the auditory and the visual information are crucial to achieve high quality synthetic speech [7]. To achieve this, the system selects from its audiovisual speech

database original combinations of auditory and visual speech. For the research described in this paper we adapted our system for visual speech-only synthesis, based on a given auditory speech signal and its text transcript. Usually, for such visual-only synthesis a phoneme-to-viseme mapping is applied. This paper focuses on the impact of applying such a mapping on the quality of the synthetic visual speech.

2. Goal

The majority of the visual speech synthesis systems found in the literature describe the target speech in terms of visemes. The input text is first written as a phonetic sequence, which is transcribed into a viseme sequence using a phoneme-to-viseme mapping. Since the number of different visemes is lower than the number of phonemes in a language, it is easier to cover all possible di-visemes, tri-visemes, etc. using a small speech database. For our experiments in English, we use the LIPS2008 [8] database, a dataset containing about 20 minutes of continuous audiovisual speech from a single speaker. For unit selection auditory or audiovisual speech synthesis, this dataset is too limited to attain high quality results. However, by using a phoneme-to-viseme mapping, the number of visual candidate units that match a target speech segment will increase, which will improve the attainable visual speech synthesis quality. On the other hand, a phoneme-based visual synthesis will always be superior to a viseme-based approach in terms of audiovisual coherence between the synthetic visual speech and the given auditory speech. Since our previous work has indicated that this coherence is crucial for a high quality perception of the audiovisual output, in this research we explore the balancing between data availability (viseme-based) and an improved audiovisual coherence (phoneme-based) for synthesis optimization.

3. Phoneme-to-viseme mapping

3.1. Standardized mapping

Various phoneme-to-viseme mappings are described in the literature. These mappings partly overlap each other, but the classification of some phonemes and the number of visemes defined tend to vary between the different implementations. In recent years, the majority of the AVTTS systems have applied a phoneme-to-viseme mapping that is described in the MPEG-4 standard [9]. This mapping uses fourteen different visemes augmented with one silence viseme.

3.2. Automatic mapping

The MPEG viseme mapping is designed to be a 'best-for-all-speakers' phoneme-to-viseme mapping. However, for usage in a data-driven AVTTS system, the phoneme-to-viseme map-

ping should be optimized for the speaker of the synthesizer’s database. To define such a speaker-dependent mapping, we first trained an AAM on the mouth-region of the frames from the database. The trained AAM consists of 8 parameters to describe the shape of the mouth (the position of the lips, chin, etc.) and 134 parameters to describe the texture of the image (the visibility of teeth, etc.). Since the AAM is trained on the actual video data from the system’s database, some of the variations that are modeled by the AAM will have nothing to do with the actual speech production, but are due to changes in the recording conditions (e.g., changes in head position in front of the camera). Therefore, we created a normalization strategy that is able to remove these variations from the model. This is achieved by fixing for every frame the value of some AAM parameters to their mean value over the database (zero). As a result, every frame of our audiovisual database is represented by 142 (8 shape and 134 texture) parameters, of which 36 (1 shape and 35 texture) parameters are normalized to zero. For more details on the AAM building and the AAM normalization technique, the reader is referred to [5].

To create the phoneme-to-viseme mapping table, for every different phoneme present in the database all its instances were gathered. Then, the normalized AAM parameters of the frame located at the middle of each instance were sampled. Afterwards, all different measures for a certain phoneme were used to create the speaker-specific mean visual representation of that phoneme. A hierarchical clustering analysis on the AAM parameters of these mean representations was performed to determine which phonemes are visibly similar for the speaker’s speaking style. Using the dendrogram, a tree diagram that illustrates the arrangement of the clusters produced by the clustering algorithm, we could discern five important steps in the hierarchical clustering procedure. Consequently, five different phoneme-to-viseme mappings were defined, using 7, 9, 11, 19 and 22 visemes, respectively. Each of these viseme sets contains a ‘silence’ viseme on which only the silence phoneme is mapped. Note that we also applied a similar procedure with multiple frames of each phoneme instance being used. However, we found that it was mainly the frame at the middle that displays the viseme’s characteristics, since the other sampled frames were often too much affected by the visual co-articulation effect.

4. Synthesis

The strategy applied for the video synthesis is very similar to the audiovisual synthesis approach described in [5]. To create a new unseen visual speech track, the system searches in the database for the most suitable sequence of video segments. As in standard unit-selection, the segments are selected based on target costs and join costs. An important target cost $C_{tar}^{context}$ is defined to take the visual co-articulation effect into account, by comparing the viseme context of the target with the viseme context of the candidate segment. To calculate this target cost, a difference matrix is constructed which expresses the similarity between every two different visemes present in the database. It is important that this matrix is calculated for the particular database that is used for synthesis, since some co-articulation effects can behave speaker-specific. For every different viseme, all its instances in the database are gathered. For each instance, the AAM parameters of the video frame located at the middle of the viseme are sampled. From these values, means M_{ij} and variances S_{ij} are calculated, where index i corresponds to the different visemes and index j corresponds to the different model

parameters. For a certain viseme i , the sum of all the variances of the different model parameters $\sum_j S_{ij}$ expresses how much the visual appearance of that viseme is affected by the visual co-articulation. Two visemes can be considered similar in terms of visual representation if their mean representations are alike and, in addition, if these mean representations are sufficiently reliable (i.e. if small summed variances were measured for these visemes). Thus, two matrices are calculated, which express for each pair of visemes the difference between their mean representations and the sum of the variances of their visual representation, respectively:

$$\begin{aligned} D_{pq}^M &= \sqrt{\sum_j (M_{pj} - M_{qj})^2} \\ D_{pq}^S &= \sum_j S_{pj} + \sum_j S_{qj} \end{aligned} \quad (1)$$

Dividing each matrix by its largest element produces the scaled matrices $D_{pq}^{M'}$ and $D_{pq}^{S'}$, after which the final difference matrix D can be constructed:

$$D_{pq} = 2 \times D_{pq}^{M'} + D_{pq}^{S'} \quad (2)$$

Matrix D can be applied to calculate the target cost of a certain unit u , matching the target viseme sequence t : the three visemes located before ($u+n$) and after ($u-n$) the unit u in the database (i.e., the unit’s viseme context) are compared to the target viseme context:

$$C_{tar}^{context} = \sum_{n=1}^3 (D(t-n, u-n)) + \sum_{n=1}^3 (D(t+n, u+n)) \quad (3)$$

Eq.3 is further optimized by adding a triangular weighting to the sum. In addition, the system uses a second target cost C_{tar}^{time} , which expresses the difference in timing between the segment and the target. The target timing is the one from the auditory speech track that is given as input to the system and the value C_{tar}^{time} increases when the synchronization of the selected segment with the target would require a heavier time-scaling of the video. To ensure smooth concatenations, a segments’s selection is also based on a join cost. The total join cost given to a certain unit is:

$$C_{join} = c_{aam} + c_{\Delta aam} \quad (4)$$

with c_{aam} the Euclidean distance between the AAM parameters of the video frames at the join position and $c_{\Delta aam}$ the Euclidean distance between the delta-AAM parameters of these frames. The total selection cost is calculated as the weighted sum of all sub-costs:

$$C_{total} = w_1 C_{tar}^{context} + w_2 C_{tar}^{time} + w_3 C_{join} \quad (5)$$

The weights $w_1-w_2-w_3$ were optimized manually.

After selection, the video segments are concatenated in order to construct the final video track. This is achieved by joining the selected AAM sub-trajectories, where the joins are further optimized by overlapping and tweaking the signals in the vicinity of the join points. More details on this are given in [5]. Finally, the joined AAM trajectories are dynamically time-scaled in order to achieve synchrony with the given auditory speech. Based on these time-scaled trajectories, inverse AAM projection leads to a new sequence of images containing the mouth-area of the talking head. Joining this video with a background video showing the other parts of the face completes the speech synthesis. Examples of the audiovisual synthesis can be found at <http://www.etro.vub.ac.be/Research/DSSP/DEMO/AVTTS>.

5. Evaluation

5.1. Objective Testing

To evaluate the quality of the different syntheses, 50 original sentences from the database were re-synthesized. Both viseme-based synthesis using automatically determined viseme groups or using the MPEG visemes and phoneme-based video selection were applied. The synthesis parameters (selection costs, etc.) were constant for all strategies. A reference synthesis strategy was added, for which phoneme-based synthesis is applied but where the video segments are selected using selection costs that are based on properties of the accompanying database audio (pitch, MFCC, etc.). In theory, this should lead to non-optimal results, since in this case no video properties are taken into account. The different strategies are summarized in table 1.

Table 1: *Synthesis strategies*

code	description
7	7 automatically determined visemes
9	9 automatically determined visemes
11	11 automatically determined visemes
17	17 automatically determined visemes
22	22 automatically determined visemes
phonv	phoneme based synthesis, video costs
phona	phoneme based synthesis, audio costs
15	15 visemes defined in MPEG

For every synthesis, the target original sentence was excluded from the database. The original database transcript was used as text input. The original database audio was used as auditory output speech and its segmentation was used as target timing. An objective quality measure was defined by comparing the synthesized AAM trajectories with the original AAM trajectories of the database video. Writing the AAM shape parameters of frame i of an original movie as \mathbf{bs}_i^{ori} and the AAM shape parameters of the same frame of the synchronized synthetic speech as \mathbf{bs}_i^{syn} , Eq.6 defines an error vector whose size is equal to the number of shape parameters:

$$e^{shape}(p) = \frac{\sqrt{\sum_{i=1}^N (bs(p)_i^{ori} - bs(p)_i^{syn})^2}}{N} \quad (6)$$

with N the number of frames in the movie. Similar as was explained in section 3.2, this vector is normalized by keeping only the speech-related parameters, resulting in vector $e_{norm}^{shape}(p)$. A similar calculation can be done for the AAM texture parameters, resulting in vector $e_{norm}^{text}(q)$. Finally, a single error value for the movie can be found:

$$E = \frac{\sum_{p=1}^{pmax} e_{norm}^{shape}(p)}{pmax} + \frac{\sum_{q=1}^{qmax} e_{norm}^{text}(q)}{qmax} \quad (7)$$

with $pmax$ and $qmax$ the number of shape and texture parameters used after normalization, respectively. Error measure E was calculated for each of the 50 movies. In figure 1 the test results obtained and the average number of candidates that were found for each segment are shown. Table 2 shows the significant error measure differences calculated by a paired t-test. Figure 1 shows that the viseme grouping definitely increases the number of candidates that is available for selection. It also indicates that this effect is less pronounced for the MPEG visemes in comparison to the automatic viseme mappings. This can be explained

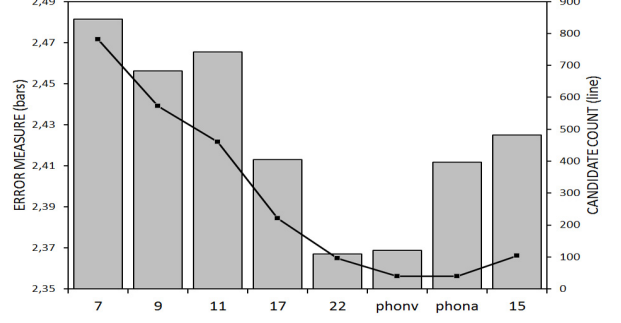


Figure 1: *Averaged error measures E (bars) and averaged number of candidates (line) for each synthesis strategy.*

Table 2: *Significant error measure differences. 'X' means significance to the 0.1 level, 'XX' to the 0.05 level.*

	7	9	11	17	22	phonv	phona	15
7	O			XX	XX	XX	XX	X
9		O		X	XX	XX		
11			O	XX	XX	XX	X	
17	XX	X	XX	O		X		
22	XX	XX	XX		O			X
phonv	XX	XX	XX	X		O	X	XX
phona	XX		X			X	O	
15	X				X	XX		O

by the fact that the automatic mapping tables group more frequently used phonemes together compared to the MPEG mapping. Surprisingly, figure 1 and table 2 indicate that the highest quality is achieved by the phoneme-based synthesis. Except for the '22' group, every viseme mapping seems to decrease the synthesis quality. If we compare the results obtained using the MPEG viseme mapping and the results obtained using the automatic viseme mappings, no significant differences were found. Also notice that even the 'phona' synthesis performs slightly better than the MPEG visemes based synthesis. Apparently, selecting the segments in a phoneme-correct fashion increases the synthesis accuracy more than the availability of a larger number of candidate segments. By comparing the 'phona' and 'phonv' results, we can see that the use of suitable selection costs further improves the synthesis.

5.2. Subjective Testing

Additionally, a subjective perception test was performed to assess the quality of the different syntheses. Based on the results of the objective test, samples from groups '9', '15', '22' and 'phonv' were chosen as test samples. We added a reference sample type 'ori', for which the original AAM trajectories from the database were used to resynthesize the visual speech. The samples were shown pairwise to the participants. Six different comparisons were considered, as described in table 3. The sequence of the comparison types as well as the sequence of the sample types within each pair were maximally randomized. The test consisted of 50 sample pairs: 14 comparisons containing an 'ori' sample and 36 comparisons between two actual syntheses. The same sentences were used for each comparison group. 13 people (aged 22-59, 9 male, 4 female) participated in the experiment, 8 of them can be considered as speech expert. The par-

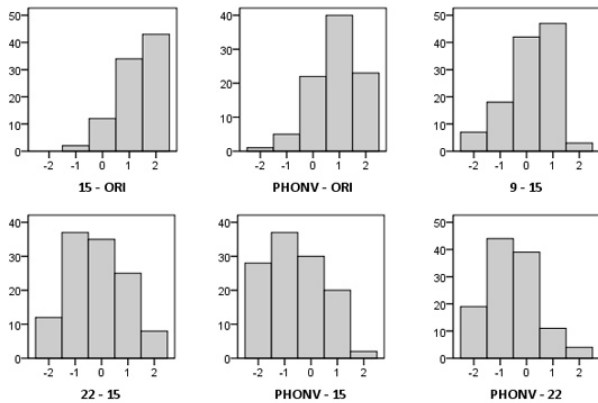


Figure 2: Subjective test results

Participants were asked to give their preference for one of the two samples of each pair using a 5-point comparative MOS scale [-2,2]. They were instructed to answer '0' if they had no clear preference. The test instructions told the participants to pay attention to both the naturalness of the mouth movements and to how well these movements are in coherence with the auditory speech that is played along with the video. The key question we asked them was: "How much are you convinced that the person you see in the sample actually produces the audio that you hear in the sample?" The results of the test are summarized in table 3 and figure 2. The observed differences for each comparison type were analyzed using a paired-sample Wilcoxon signed rank test. These test results are also given in table 3.

Table 3: Subjective evaluation

Type 1	Type 2	N	Mean	Z	Sign.
15	ori	91	1.3	-7.79	< 0.01
phonv	ori	91	0.87	-6.46	< 0.01
9	15	117	0.18	-1.94	0.052
22	15	117	-0.17	-1.64	0.102
phonv	15	117	-0.59	-5.24	< 0.01
phonv	22	117	-0.54	-5.04	< 0.01

The results observed for the subjective test are in line with the results of the objective test described in section 5.1. The participants were clearly in favor of the synthesis based on phonemes compared to the viseme-based syntheses. Furthermore, a perception of higher quality is attained by increasing the number of visemes. The synthesized samples are still distinguishable from natural visual speech, although also for this aspect the phoneme-based syntheses outperforms the viseme-based approaches.

6. Conclusions

In this paper we have studied the use of visemes in (audio-) visual speech synthesis. For this, we adapted our audiovisual speech synthesis system to synthesize visual speech only, based on viseme labels. In theory, such a viseme-based unit selection should outperform the phoneme-based selection since it multiplies the number of candidate units for selection. Nevertheless, in both objective and subjective evaluations a synthesis based on phonemes resulted in better ratings compared to the

syntheses based on visemes. In addition, the results obtained show that the synthesis quality increases when more different visemes are considered. We also found that the standardized MPEG-4 viseme set performs comparable to the automatically determined viseme sets. These results raise some questions on the viseme-based approach that is widely applied in (audio-) visual speech synthesis. It appears that the precise description of each visual speech element in a phoneme-based synthesis leads to a more accurate prediction of the synthetic speech. In addition, when the synthetic visual speech is displayed in an audio-visual manner to the observer, the enhanced multimodal coherence of the phoneme-based synthesis increases the perception quality. This result is in line with our previous results, where we concluded that this audiovisual coherence is one of the main determinants for a high-quality perception. On the other hand, we believe that the many-to-one phoneme-to-viseme mappings that have been used in this research insufficiently describe all the fine details of the visual speech information. Although the synthesizer mimicked the visual co-articulation effect by applying a target cost based on the viseme-context, we believe that better viseme-based synthesis results can be achieved by implementing a many-to-many phoneme-to-viseme mapping that describes the visual co-articulation effect as well. Using such a viseme set, it should be interesting to re-evaluate the balancing between available data (viseme-based) and an improved audio-visual coherence (phoneme-based) for synthesis optimization.

7. Acknowledgments

Parts of the research reported on in this paper were performed in the context of the project CAdeE: Toward Cognitive Adaptive Edugames (HOA26). The authors would like to thank the participants of the subjective perception test for their time.

8. References

- [1] Chen, T., "Audiovisual speech processing", IEEE Signal Processing Magazine, 8:9-21, 2001.
- [2] Hilder, S., Theobald B.J., Harvey, R., "In Pursuit of Visemes", International Conference on Auditory-Visual Speech Processing (AVSP), 154-159, 2010.
- [3] Jackson, P.L., "The theoretical minimal unit for visual speech perception: visemes and coarticulation", Volta Review, 90(5):99-115, 1988.
- [4] Ezzat, T., Poggio, T., "Visual speech synthesis by morphing visemes", International Journal of Computer Vision, 38(1):45-57, 2000.
- [5] Mattheyses W., Latacz L., Verhelst W., "Optimized Photorealistic Audiovisual Speech Synthesis Using Active Appearance Modeling", International Conference on Auditory-visual Speech Processing '10, 2010.
- [6] Edwards, G.J., Taylor, C.J., Cootes, T.F., "Interpreting Face Images using Active Appearance Models", Int. Conf. on Face and Gesture Recognition, 300-305, 1998.
- [7] Mattheyses, W., Latacz, L. and Verhelst, W., "On the importance of audiovisual coherence for the perceived quality of synthesized visual speech", EURASIP Journal on Audio, Speech, and Music Processing, SI: Animating Virtual Speakers or Singers from Audio: Lip-Synching Facial Animation, 2009.
- [8] Theobald, B.-J., Fagel, S., Bailly, G., Elisei, F., "LIPS2008: Visual speech synthesis challenge", Interspeech '08, 1875-1878, 2008.
- [9] Pandzic, I.S., Forchheimer, R., "MPEG-4 Facial Animation: The Standard, Implementation and Applications", John Wiley & Sons, Inc., 2003.