ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
(AVSP) 2013
Annecy, France
August 29 - September 1, 2013

Visual Control of Hidden-Semi-Markov-Model based Acoustic Speech Synthesis

Jakob Hollenstein[1,2],*Michael Pucher*[1], *Dietmar Schabus*[1,3]

[1]Telecommunications Research Center Vienna (FTW), Vienna, Austria
[2]Vienna University of Technology, Vienna, Austria
[3]Graz University of Technology, Graz, Austria
{hollenstein,pucher,schabus}@ftw.at

Abstract

We show how to visually control acoustic speech synthesis by modelling the dependency between visual and acoustic parameters within the Hidden-Semi-Markov-Model (HSMM) based speech synthesis framework. A joint audio-visual model is trained with 3D facial marker trajectories as visual features. Since the dependencies of acoustic features on visual features are only present for certain phones, we implemented a model where dependencies are estimated for a set of vowels only. A subjective evaluation consisting of a vowel identification task showed that we can transform some vowel trajectories in a phonetically meaningful way by controlling the visual parameters in PCA space. These visual parameters can also be interpreted as fundamental visual speech motion components, which leads to an intuitive control model.

Index Terms: audio-visual speech synthesis, HMM-based speech synthesis, controllability

1. Introduction

One key strength of the HSMM-based speech synthesis framework [1] lies in its greater flexibility in comparison to waveform concatenation methods, often accredited to the possibility to use model adaptation [2] and interpolation [3]. In addition to these data-driven approaches to diversify the characteristics of synthetic speech, methods that allow more direct control using phonetic background knowledge have been proposed more recently. Acoustic speech characteristics have been successfully modified by exercising control on articulartory [4] as well as on formant [5] parameters. This is achieved by training piecewise linear transformations from the models for the articulatory or formant domain to the models for the acoustic domain, using a multimodal data corpus. Similar to these works, in this paper we investigate the possibility of using visual speech features based on facial marker motion data to modify and control acoustic synthetic speech. Our work is similar to [4], but uses more restricted features (e.g., no tongue positions) which are easier to record.

Possible use cases of this include more intuitive control of speech synthesis, the possibility to use physically intuitive data to constrain trajectories as well as the possibility to use this information in language learning to provide clues of required changes.

To investigate the possibility of visual control, we modified the system used in [5] to control acoustic speech synthesis by visual features (instead of formants). The same line spectral pairs features as in [4] are adopted as acoustic features in our approach.

2. Data and System

We work with a corpus of synchronous audio and facial motion recordings [6] where facial motion was recorded using an OptiTrack system [7], which records the 3D positions of 37 markers glued to a speaker's face at 100 Hz. This corpus was originally recorded for speaker-adaptive audio-visual speech synthesis [8]. In this paper, we used one male speaker's data consisting of 223 Austrian German sentences amounting to roughly 11 minutes total.

2.1. Training

In a first attempt, we replaced the formant stream in the system described in [5] with a visual stream, resulting in a joint audio-visual model. The visual features were computed from the 3D facial marker coordinates via Principal Component Analysis (PCA) as described in [6]. However, by comparing the variation of the spectral features with respect to the different phones to the variation induced by the transformation when modifying PCA features, we found that the changes due to the transformations were very small. Hence it would be impossible to achieve a natural variety of different phones in this way. A further reduction of the number of visual dimensions lead to an even smaller amount of change induced by the transformations, hinting at the insufficient explanation of the spectral features by the visual features.

To improve the expressiveness of the features with respect to different phones and increase the ability to interpret them, PCA was abandoned and raw coordinate features used. To reduce the dimensionality of the raw visual features, a selection of some visual markers with a direct influence on speech production was made: mouth opening and lip protrusion, represented by the markers *jaw, lower lip* and *upper lip*. This is similar to the facial markers in [9]. The movement of these markers in the left/right direction was assumed to be negligible, thus only the Y and Z axes were used. A new joint audio-visual model was trained with these features. To ease visualization PCA was done on this restricted model.

It is not possible to distinguish all phones merely by their visual features. Depending on the speaker, the visual features overlap considerably for different phones, since the visual features mostly capture openness and roundedness. This explains the initial results and is similiar to the reason for using visemes in visual speech synthesis[10].

Figure 1 shows the visual features retrieved from forced alignment of the training data with respect to all vowels and diphthongs. A bagplot [11] is used to illustrate the distribution of the naturally occurring trajectories. This is done in

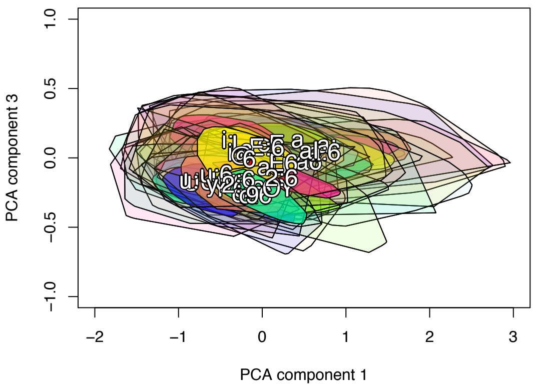

Figure 1: *Bagplot showing the distribution of vowels and diphthongs from the training data in PCA1×PCA3 space.*

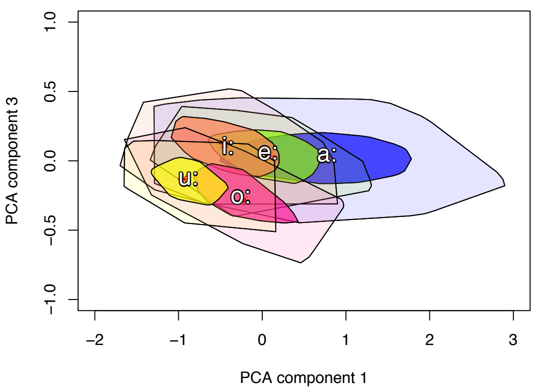

Figure 2: *Bagplot showing the distribution of selected vowels (*/a: e: i: o: u:/*) from the training data in PCA1×PCA3 space.*

PCA1×PCA3 space.

Since the visual data for different phones overlaps, we decided to constrain the dependency modeling on a small set of vowels.

The selection of phones that contribute to a transformation is a trade-off between a set comprising more and a wider variety of phones and more distinctive visual representations. For our experiments the set of vowels /a: e: i: o: u:/ was chosen. Figure 2 shows the grouping and overlap of the visual features in PCA1×PCA3 space. Notice how this resembles the vowel trapezium (rotated and mirrored, open-close from right to left). While the PCA1×PCA2 space showed even more resemblance with the vowel diagram, the PCA1×PCA3 space provides more distinction between /o: u:/ and /a: e: i:/. Which is consistent with better control regarding changes from /o: u:/ to /a: e: i:/ and vice versa and also the reason for choosing PCA1×PCA3 instead of the former.

The system uses a common clustering for acoustic and visual features, and thus for each acoustic leaf there is a corresponding visual leaf. Each leaf in the clustering tree can be assigned a transform and each transform can be assigned to several leaves. We used a single global transform for the selected vowels. To achieve this clustering, we introduced an /a: e: i: o: u:/ question at the root of the clustering tree. For all models outside of the /a: e: i: o: u:/ subtree, no transformations were trained, as illustrated in Figure 3.

Since we adapted the system from [5], we applied the same

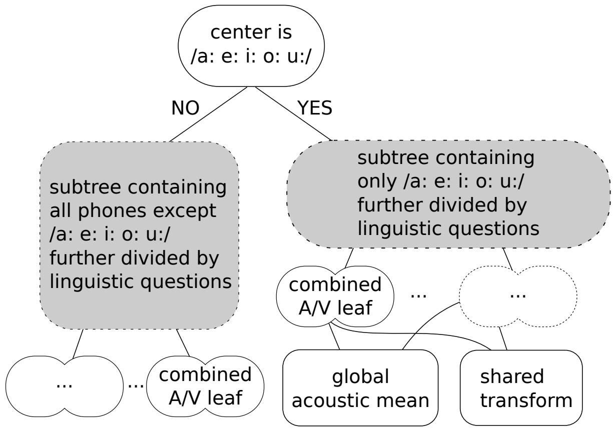

Figure 3: *Clustering tree with* /a: e: i: o: u:/ *central phone question.*

equations, described in more detail in [9], for EM estimation of the visual means $\mu_{\mathbf{Y}_j}$:

$$\mu_{\mathbf{Y}_j} = \frac{\sum_{t=1}^{T} \gamma_j(t)\mathbf{y}_t}{\sum_{t=1}^{T} \gamma_j(t)}, \tag{1}$$

where $\mathbf{y}_t$ and $\mathbf{x}_t$ describe the visual and acoustic observation vectors at time t respectively, γ_j describes the state occupancy probability of state j and T is the total length of training data.

For the acoustic means $\mu_{\hat{\mathbf{X}}_j}$ of the models in the /a: e: i: o: u:/ subtree, we use the dependency model parameters to apply a linear transformation from visual to acoustic parameters,

$$\mu_{\hat{\mathbf{X}}_j} = \frac{\sum_{t=1}^{T} \gamma_j(t)(\mathbf{x}_t - \hat{\mathbf{A}}_j\mathbf{y}_t)}{\sum_{t=1}^{T} \gamma_j(t)}. \tag{2}$$

$$\hat{\mathbf{A}}_j = \frac{\sum_{t=1}^{T} \gamma_j(t)(\mathbf{x}_t - \mu_{\hat{\mathbf{X}}_j})\mathbf{y}_t^T}{\sum_{t=1}^{T} \gamma_j(t)\mathbf{y}_t\mathbf{y}_t^T}. \tag{3}$$

The estimation of the linear transformations $\hat{\mathbf{A}}_j$ is also constrained to models in the /a: e: i: o: u:/ subtree. For the acoustic means $\mu_{\mathbf{X}_j}$ of all other phones, as well as for all (acoustic and visual) variances, we used the un-transformed version as in Equation (1).

In contrast to [9], the mean of all acoustic leafs sharing the same transform is also shared in our implementation. This is done by employing the tying mechanisms in HTS/HTK. Thus different states still share the same underlying $\hat{\mathbf{A}}_j$ regardless of the state j. The use of the tying mechanism is explained in [12].

Thus the transformation of the visual features to the audio features is not used to superimpose the modified trajectory on the original audio feature trajectory but is effectively used to generate the audio feature trajectory from the visual feature trajectory. This means that without visual information, all /a: e: i: o: u:/ phones would result in the same acoustic realization. Using this approach, we implemented a constrained audio-visual dependency modeling system.

2.2. Parameter Generation

For parameter generation, we implemented a simplified version of the algorithm described in [9]. Given an optimal state sequence, the optimal acoustic parameter sequence X_S^* is generated as

$$\mathbf{X}_S^* = (\mathbf{W}_\mathbf{X}^T \mathbf{U}_\mathbf{X}^{-1} \mathbf{W}_\mathbf{X})^{-1} \mathbf{W}_\mathbf{X}^T U_\mathbf{X}^{-1} (\mathbf{M}_\mathbf{X} + \mathbf{A} \mathbf{W}_\mathbf{Y} \mathbf{Y}_S) \quad (4)$$

which results from

$$\frac{\partial \log P(\mathbf{W}_\mathbf{X} \mathbf{X}_S, \mathbf{W}_\mathbf{Y} \mathbf{Y}_S | \lambda, q^*)}{\partial \mathbf{X}_S} = 0. \quad (5)$$

The visual parameter sequences $\mathbf{Y}_S^*$ are generated based on the approximation

$$\frac{\partial \log P(\mathbf{W}_\mathbf{Y} \mathbf{Y}_S | \lambda, q^*)}{\partial \mathbf{Y}_S} \approx \frac{\partial \log P(\mathbf{W}_\mathbf{X} \mathbf{X}_S, \mathbf{W}_\mathbf{Y} \mathbf{Y}_S | \lambda, q^*)}{\partial \mathbf{Y}_S} \quad (6)$$

which results in

$$\mathbf{Y}_S^* = (\mathbf{W}_\mathbf{Y}^T \mathbf{U}_\mathbf{Y}^{-1} \mathbf{W}_\mathbf{Y})^{-1} \mathbf{W}_\mathbf{Y}^T U_\mathbf{Y}^{-1} \mathbf{M}_\mathbf{Y} \quad (7)$$

when we set the left hand side of Equation (6) to 0. Note that this is the standard parameter generation algorithm described in [13].

3. Visual Control using PCA Features

To simplify the control from having to modify points in 6 dimensional space, we apply PCA. To modify a given model, the means are transformed into PCA space, modifications are performed relative to the resulting PCA feature vector, and the modified vector is projected back into the original space. No dimensionality reduction is used in this scheme. Also, the trajectory is not modified directly (e.g., by adding to the trajectory values), but the means are changed, thus changing the generated trajectory. This ensures smooth trajectories for the duration of the modified phone and especially at the phone boundaries. As a side effect of the smoothing, the extent of modification is slightly decreased, thus a larger change in the control parameters is required to achieve sufficiently strong effects.

The first PCA component roughly corresponds to mouth opening, while the second and third component can be interpreted as modelling rounding. Figure 4 illustrates the changes of the marker positions resulting from changes of the first PCA component between -1.5 and $+1.5$ and between -0.75 and $+0.75$ for the second PCA component.

We did not carry out a formal evaluation of the effects of the control on the visual speech motion, but synthesis of the entire 37 marker positions can be performed at the loss of some accuracy by calculating a linear regression from the 6 visual control parameters to the full visual parameter space. From examples we looked at during development, it appeared that visual synthesis is still feasible using only these parameters. There is also some loss of acoustic quality due to the simple transformation and the incomplete explanation of acoustic features by the visual features.

Figure 5 illustrates an example outcome. In the sentence *“Ich habe ‘bomo’ gehört”* (I heard ‘bomo’), the two vowels of the nonsense word ‘bomo’ were modified visually by increasing and decreasing the mean of the first PCA component, corresponding to increased and decreased mouth opening, respectively. The bottom part of the figure shows the effect on the distance between the upper lip and the lower lip markers. The middle part shows the resulting spectrograms for the time segment indicated by vertical lines. The top part of the figure shows the resulting facial marker configurations at the time points indicated by small circles. Compared to the unmodified sample (± 0, center spectrogram, first formant at 287 Hz), the samples with the decreased (-1.5, left spectrogram, first formant

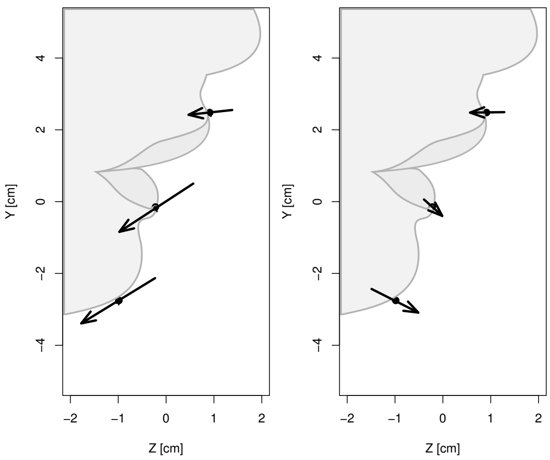

Figure 4: *Effects of changing the first PCA component from -1.5 to $+1.5$ (left) and changing the third PCA component from -0.75 to $+0.75$ (right).*

at 408 Hz) and increased ($+1.5$, right spectrogram, first formant at 605 Hz) mouth opening exhibit a clearly visible change both in the visual trajectories as well as in the spectral power distributions.

4. Evaluation

To evaluate the acoustic effects of the visual control, a subjective listening test with eleven subjects was carried out. We synthesized ten utterances containing the nonsense words ‘bama’, ‘beme’, ‘bimi’, ‘bomo’, ‘bumu’ and ‘pata’, ‘pete’, ‘piti’, ‘poto’, ‘putu’ in the carrier sentence *“Ich habe … gehört”*, with varying visual control parameters affecting the two vowels of the nonsense word. The first PCA component was modified by applying one of the three offsets $-1.5, \pm 0, +1.5$ and the third PCA component by applying one of the three offsets $-0.75, \pm 0, +0.75$, resulting in nine different realizations[1].

Each test subject heard all 90 synthesized examples in random order and was asked to identify what they heard as one of the five variants of the nonsense word (either ‘bama’, ‘beme’, ‘bimi’, ‘bomo’, ‘bumu’ or ‘pata’, ‘pete’, ‘piti’, ‘poto’, ‘putu’) or none of these. The results are given in Table 1, where the first column gives the original vowel, the second and third column give the applied offsets for the first and third PCA components, and the remaining columns give the identification percentages. These results are also visualized in Figure 6 as stacked bar charts. For each initial vowel, the central bar corresponds to the unmodified sample, the upper-left bar corresponds to a shift in the upper-left direction (compare Figure 2), etc.

In most cases, there is a clear majority regarding the perceived vowel and in these cases the change is consistent with what can be expected when we consider the vowel distributions of Figure 2. For each initial vowel, we can successfully transform towards at least one other vowel. It is interesting that for each of them there is a direction in which the listeners perceived the original vowel more clearly, i.e. in a higher percentage of cases. In all five cases, this direction is leading “away” from the other vowels (Figure 2). Furthermore, we see that for /a/, there is a fairly large number of “no match” votes (“?”) in all nine cases. Part of this is due to acoustic artifacts (e.g., buzzing

[1] Examples on http://userver.ftw.at/~schabus/avsp2013vc

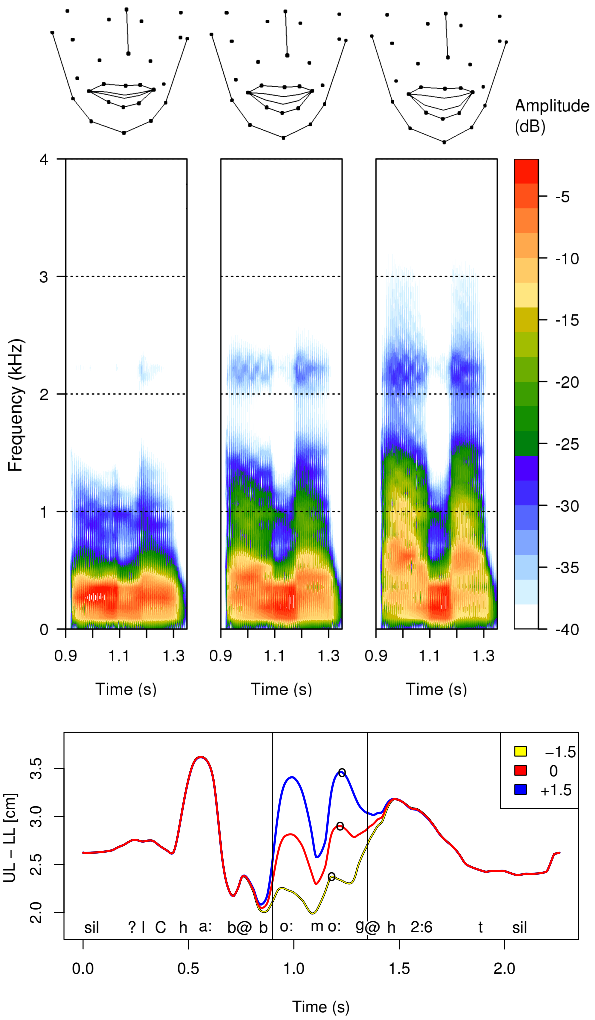

Figure 5: *Example outcome of modification of the first PCA component. Bottom: Distance between the upper lip and lower lip markers over time. Middle: spectrograms for the time segment indicated by vertical lines. Top: Facial marker configuration at the time points indicated by small circles.*

or distortions regarding amplitude). These artifacts can be attributed to leaving the area of the PCA space in which observations are naturally occurring and thus creating artificial visual features and inducing artificial sound.

In some examples (e.g., the one in Figure 5) we saw that the closure between the two modified vowels was not synthesized correctly in the visual domain as the smooth trajectory generation prevented the lip movement from reaching the closure point. This could be prevented, e.g., by decreasing the variances and thus forcing the trajectories closer to the specified means, or by modifying the dynamic features. This also indicates that PCA transformation alone does not capture all possible modifications of the underlying feature space adequately and a different parametrization for exercising control may be necessary.

We saw in our experiments that the offsets applied in PCA space needed to be larger than expected (i.e., 1.5 and 0.75 instead of 1.0 and 0.5) in order to properly induce changes. We attribute part of this to the generation algorithm producing over-

Table 1: Evaluation Results: Identification percentages for each initial vowel, modified by each of nine control offset combinations.

V	Δ_1	Δ_3	a	e	i	o	u	?
a	-1.5	+0.75	0	77.3	0	0	0	22.7
a	0	+0.75	4.5	72.7	0	0	0	22.7
a	+1.5	+0.75	9.1	40.9	0	0	0	50.0
a	-1.5	0	22.7	36.4	0	0	0	40.9
a	0	0	63.6	0	0	0	0	36.4
a	+1.5	0	72.7	0	0	0	0	27.3
a	-1.5	-0.75	0	0	0	68.2	0	31.8
a	0	-0.75	4.5	0	0	54.5	0	40.9
a	+1.5	-0.75	36.4	0	0	18.2	0	45.5
e	-1.5	+0.75	0	27.3	72.7	0	0	0
e	0	+0.75	0	100.0	0	0	0	0
e	+1.5	+0.75	0	100.0	0	0	0	0
e	-1.5	0	0	0	90.9	0	9.1	0
e	0	0	0	68.2	22.7	0	0	9.1
e	+1.5	0	4.5	95.5	0	0	0	0
e	-1.5	-0.75	0	0	45.5	0	50.0	4.5
e	0	-0.75	0	0	13.6	0	72.7	13.6
e	+1.5	-0.75	13.6	31.8	0	31.8	18.2	4.5
i	-1.5	+0.75	0	4.5	86.4	0	9.1	0
i	0	+0.75	0	95.5	4.5	0	0	0
i	+1.5	+0.75	0	95.5	4.5	0	0	0
i	-1.5	0	0	0	95.5	0	4.5	0
i	0	0	0	31.8	50.0	4.5	13.6	0
i	+1.5	0	0	81.8	4.5	9.1	4.5	0
i	-1.5	-0.75	0	0	54.5	0	40.9	4.5
i	0	-0.75	0	0	40.9	0	50.0	9.1
i	+1.5	-0.75	0	50.0	0	27.3	18.2	4.5
o	-1.5	+0.75	0	63.6	13.6	0	13.6	9.1
o	0	+0.75	0	86.4	0	0	0	13.6
o	+1.5	+0.75	4.5	86.4	0	0	0	9.1
o	-1.5	0	0	0	4.5	0	90.9	4.5
o	0	0	0	18.2	0	72.7	9.1	0
o	+1.5	0	72.7	9.1	0	13.6	0	4.5
o	-1.5	-0.75	0	0	0	0	100.0	0
o	0	-0.75	0	0	0	27.3	72.7	0
o	+1.5	-0.75	4.5	0	0	90.9	0	4.5
u	-1.5	+0.75	0	0	95.5	0	4.5	0
u	0	+0.75	0	22.7	68.2	0	0	9.1
u	+1.5	+0.75	0	81.8	13.6	0	0	4.5
u	-1.5	0	0	0	27.3	0	72.7	0
u	0	0	0	0	22.7	0	63.6	13.6
u	+1.5	0	0	27.3	4.5	9.1	50.0	9.1
u	-1.5	-0.75	0	0	0	0	90.9	9.1
u	0	-0.75	0	0	0	0	90.9	9.1
u	+1.5	-0.75	0	0	0	0	95.5	4.5

smoothed trajectories and part of it to the overlapping which essentially requires us to leave ambiguous areas to create unambiguous sounds.

5. Conclusion

In this paper we have shown that – similar to previous work, where acoustic speech synthesis is controlled via articulatory or formant features – it is also possible to achieve phonetically meaningful transformations of acoustic synthetic speech by exercising control in terms of visual speech features, namely 3D facial marker motion data. This can be seen as a more restrictive setting than 3D articulatory data, because we have no information on the position of the tongue.

For all of the selected five phones, transformations to at least one other phone have been shown to be feasible, as determined by a subjective listening test with eleven subjects. The acoustic phone realizations resulting from changes in certain directions are consistent with the distribution in visual PCA space.

Future work could explore the improvements achievable by using a more sophisticated control model for example using combined per-state and per-context transformations. Some improvement may also be possible by using more descriptive

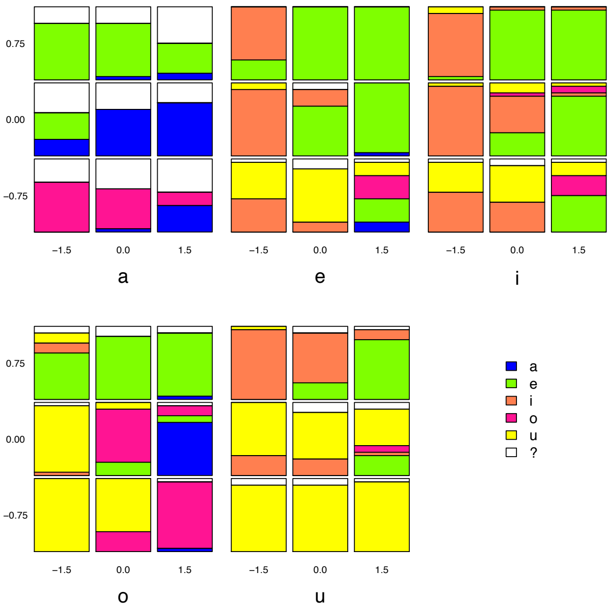

Figure 6: *Visualization of the evaluation results (Table 1). For each initial phone, the central subplot shows the classification results for the unmodified phone. The eight surrounding subplots show the classification results for the modified phones. Colors and orientation are in line with Figure 2.*

features, rather than only three markers on the lips and jaw. Additionally, we would like to evaluate the coherence of the modified visual and acoustic speech signals in combined audio-visual perceptive experiments. Another interesting topic to investigate would be to combine both facial marker and articulatory data, requiring a synchronous multimodal corpus, which we plan to build in the near future.

6. Acknowledgements

We want to thank Korin Richmond for providing a HMM-based system with dependency modeling. This work was supported by the Austrian Science Fund (FWF): P22890-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET - Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG.

7. References

[1] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. SSW*, 2007, pp. 294–299.

[2] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[3] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Proc. Eurospeech*, 1997.

[4] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 573–576.

[5] M. Lei, J. Yamagishi, K. Richmond, Z.-H. Ling, S. King, and L.-R. Dai, "Formant-controlled HMM-based speech synthesis," in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 2777–2780.

[6] D. Schabus, M. Pucher, and G. Hofer, "Building a synchronous corpus of acoustic and 3D facial marker data for adaptive audio-visual speech synthesis," in *Proc. LREC*, Istanbul, Turkey, May 2012, pp. 3313–3316.

[7] Naturalpoint, 2013. [Online]. Available: http://www.naturalpoint.com/optitrack/

[8] D. Schabus, M. Pucher, and G. Hofer, "Speaker-adaptive visual speech synthesis in the HMM-framework," in *Proc. Interspeech*, Portland, OR, USA, 2012, pp. 979–982.

[9] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *Trans. Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[10] T. Chen, "Audiovisual speech processing," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 9–21, 2001.

[11] P. J. Rousseeuw, I. Ruts, and J. W. Tukey, "The bagplot: a bivariate boxplot," *The American Statistician*, vol. 53, no. 4, pp. 382–387, 1999.

[12] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[13] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, Detroit, MI, USA, 1995, pp. 660–663.