Integration of Acoustic and Visual Cues in Prominence Perception

Hansjörg Mixdorff[1], *Angelika Hönemann*[1], *Sascha Fagel* [2]

[1] Department of Computer Science and Media, Beuth University Berlin, Germany
[2]Zoobe Message Entertainment GmbH, Berlin, Germany

[mixdorff|ahoenemann]@beuth-hochschule.de, fagel@zoobe.com

Abstract

This study concerns the perception of prominence in auditory-visual speech perception. We constructed A/V stimuli from five-syllable sentences in which every syllable was a candidate for receiving stress. All syllables were of uniform length, and the *F0* contours were manipulated using the Fujisaki model, moving a peak of *F0* from the beginning to the end of the utterance. The peak was either aligned with the center of the syllable or the boundary between syllables, yielding a total of nine positions. Likewise, a video showing the upper part of a speaker's face exhibiting one single raise of eyebrows was aligned with the audio, hence yielding nine positions for the visual cue, with the maximum displacement of the eyebrows coinciding with syllable centers or boundaries. Another series of stimuli was produced with head nods as the visual cue. In addition stimuli with constant F0 with or without video were created. 22 German native subjects rated the strength of each of the five syllables in a stimulus on a scale from 1-3. Results show that the acoustic prominence outweighs the visual one, and that the integration of both in a single syllable is the strongest when the movement as well as the *F0* peak are aligned with the center of the syllable. However, *F0* peaks aligned with the right boundary of the accented syllable, as well as visual peaks aligned with the left one also boost prominence considerably. Nods had an effect similar in magnitude as eye brow movements, however, results suggest that they rather have to be aligned with the right boundary of the syllable than the left one.

Index Terms: Prominence, auditory-visual integration, *F0* modeling

1. Introduction

It is evident that speech perception benefits from visual contact and that the two channels of communication are integrated and influence the result of perception. The famous McGurk effect shows that the two senses are strongly connected and conflicting cues are resolved to form the most likely percept [1]. The first author and his co-authors have also shown that syllabic tone perception in noise is facilitated by seeing the talker's face [2]. In more recent work the authors of the current study investigated non-verbal visual cues and their connection with speech prosody [4]. To this end, a corpus of spontaneous A/V speech was collected and annotated on the acoustic as well as the visual level. During the alignment of acoustic landmarks, such as accents and boundaries, with visible non-speech movements the question arose, in which way the anchoring of movements should be performed. In a first restricted approach, only movements occurring during accented syllables or syllables preceding a boundary were taken into account. However, this left a number of movements unanchored as they were located in syllables neighboring accented syllables, for instance. For this reason we designed the perceptual experiment reported in the current paper investigating in which way acoustic and visual cues have to be aligned to reinforce the perceived prominence of the same underlying syllable(s), and at what distance they would represent separate events of prominence. It has been shown in earlier studies on auditory-visual prominence that the acoustic usual surpasses the visual cue in strength (see, for instance, [3]), however, we were mostly interested in closely looking at how precisely the cues have to be aligned with each other to be either perceived as one event or not.

2. Stimulus Design and Experiment Procedure

We created three five-syllabic sentences of German in which each syllable was a mono-syllabic word and a candidate for being accented. We also aimed to create maximally sonorant sequences in order for the *F0* contour to be continuous:

Sentence	English
Bens Haar war sehr lang.	*Ben's hair was very long.*
Jims Rad war nie grün.	*Jim's bike was never green.*
Johns Bein war ganz blau.	*John's leg was all blue.*

These sentences were synthesized at an *F0* of 100Hz using *MBROLA* [5] and the German male voice *de8* (22050Hz, 16bit), keeping the duration of each syllable at 300ms in order to minimize the influence of durational cues on the percept of prominence. One reason for choosing the synthetic voice was to yield uniform intensities for all the syllables, as intensity is also an important correlate of perceived prominence.

The durations of the phones in each syllable were determined by segmenting natural, monotonously uttered recordings of the sentences by the first author using the *PRAAT TextGrid* Editor [6] and setting the phone durations of the stimuli in proportion to the natural syllabic durations observed.

Fujisaki model-based [7] *F0* contour parameterization [8] was performed on natural utterances uttered by the first author with a single accent placed on one of the five syllables and yielded configurations with one phrase component and one accent component. From these parameters we derived standard settings for the synthetic utterances, keeping the underlying phrase command the same for all stimuli (*Ap*=.26, start time 330ms before utterance onset) while shifting the accent command (*Aa*=.45, or the equivalent of an interval of 7 st, duration=150ms) by increments of 150ms in order for the *F0* peak to coincide with either the center or the boundary of the syllables, starting with the center of the first syllable and ending with the center of the last syllable, hence yielding nine different alignments of acoustic prominence. *Fb* was set to 92Hz. Figure 1 (left) shows an example of such an audio

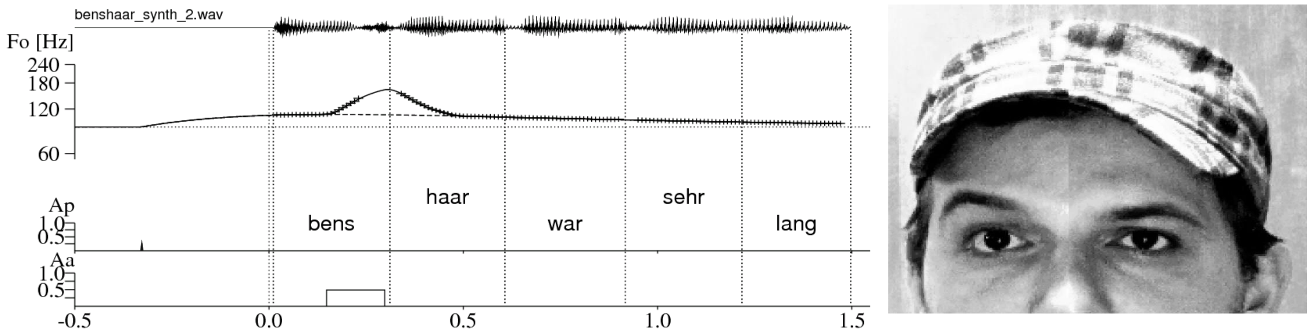

Figure 1: *Example of a stimulus (left: Audio, right: video). The audio stimulus shows an example where the F0 peak is aligned with the boundary between the first and the second syllable. The picture on the right displays the right side of the face at the moment of greatest displacement of the eye brows alongside the left side in the resting position.*

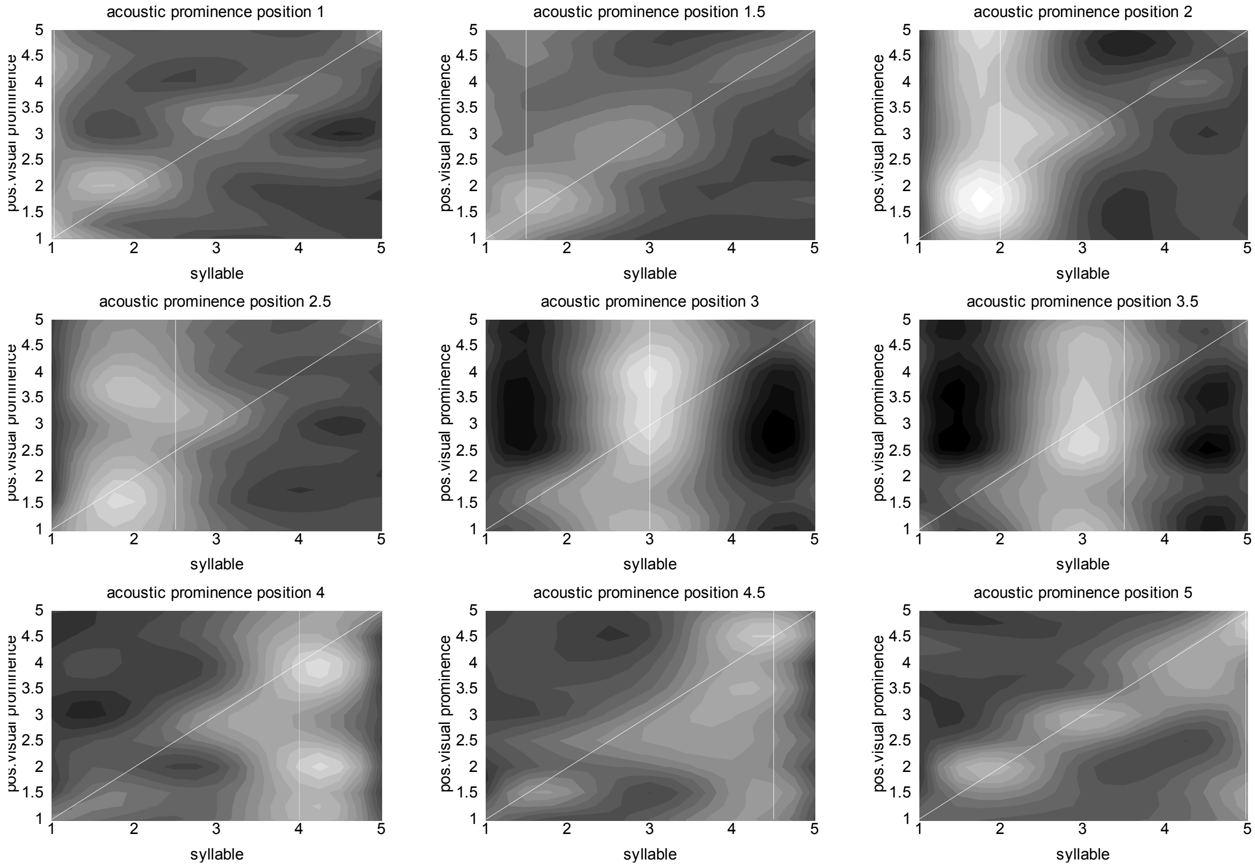

Figure 2: *Visualization of experiment results for stimuli AVEF. Brighter shading represents higher perceived prominence. The x axis represents the five syllables of the sentences, the y axis the nine alignment positions for the eye brow movement, also indicated by the diagonal white line. The vertical line in each panel indicates the position of the acoustic prominence. See text for details. Panel numbering: top row 1, 2 and 3; center row 4, 5 and 6; bottom row 7, 8 and 9.*

stimulus. The figure displays from the top to the bottom: The speech wave form, the *F0* contour (+signs: extracted, solid line: model-based), the text, the underlying phrase and accent commands. In this example the peak of *F0* was adjusted to coincide with the boundary between the first and the second syllable. All *F0* modifications were performed on the monotonous audio using the *FujiParaEditor* [9] driving the *PRAAT* PSOLA resynthesis.

The video part of the stimuli was created by asking a male subject to sit still and simply raise both of his eyebrows simultaneously from time to time without talking. The upper part of his head was filmed using a Panasonic mini-DV camera (PAL, 576i50, landscape orientation). By limiting the visual stimulus to single eye brow raises we aimed to have close control of where activity in the visual channel occurred. To that end only the upper part of the face was presented hence concealing that the subject was actually not talking. We selected a single instance of eye brow raises surrounded by inactivity. The movement lasted 11 frames in the video or 440ms. Likewise we asked the subject to produce light nods of the head. The instance that we eventually chose lasted 15 frames or 600ms. We then used the monotonous audio track to align the maximum displacement of the visual cues with either the center or the boundary of the syllables, starting again with the center of the first syllable and ending with the center of the last one, yielding nine different videos. All the video editing was performed in *Adobe Premiere* v. 6.5. After exporting the video clips they were loaded into *VirtualDub* [10], de-interlaced by duplicating the odd fields and the part of the face underneath the tip of the nose was removed as shown in Figure 1 (right), yielding a final size of 720 x 531 pixels. The maximal vertical displacement of the eye brows was about 38 pixels. The picture shows the moment of largest displacement (left) alongside the resting position (right).

For the nod movement we only created versions for the sentence "Jims Rad…" with the nod aligned to the centers of the second, third and fourth syllable, and the maximum vertical displacement of the head was 89 pixels.

Finally we combined all audio and video versions with each other, also using *VirtualDub*, yielding the following types of stimuli which we created for each of the three sentences, except for the nod that was combined with one sentence only:

	number of stimuli	stimulus type
audio only		
monotonous	1 x 3	AOM
Fujisaki model-based	9 x 3	AOF
audio+video (eye brows)		
monotonous	9 x 3	AVEM
Fujisaki model-based	81 x 3	AVEF
audio+video (nod)		
Fujisaki model-based	27 x 1	AVNF

This yielded a total of 100 x 3+27= 327 stimuli. The complete list of stimuli was randomized and manually checked in order to avoid frequent repetitions of the same sentence in a sequence. Then the randomized list was split into three sets of 109 stimuli each in order to make the task more manageable. The experiment was programmed as a desktop application. In the intro we explained that the experiment was about audio-visual speech synthesis and the ability to create subtle differences in meaning when controlling a virtual agent. Subjects were asked to closely view the image (if present) when listening to the stimulus. Then they had to decide for each of the five syllables whether it had been accented weakly (level 1), average (level 2) or strongly (level 3). In the screen for stimulus presentation they were provided the five words of the sentence and five number fields initialized with "1" which they were supposed to edit, otherwise they could not advance to the next stimulus. Two filled-in examples were presented before the beginning of the actual experiment. Subjects were allowed to listen to the stimuli as often as they wanted and after they had made their decision advanced to the next stimulus by pressing the button "Next". Playback was done using inexpensive headsets, and the experiment was performed in sound-untreated class room fitted with 20 desktop computers.

Participants were 35 students of Media Informatics at Beuth University, of these 17 male and 5 female German native listeners between 20 and 31 years of age. Each of the subjects who either had corrected or normal vision, as well as normal hearing, was assigned one of the three stimulus sets containing 109 stimuli. The experiment took between 20 and 42 minutes to complete. The distance of participants from the computer displays was not explicitly controlled. Participation was rewarded by course credits. The results presented in this paper are from the 22 German native subjects.

3. Results of Analysis

It must be stated that the three stimulus sets were assigned to students without prior knowledge of their language backgrounds. Therefore the German listeners were not equally distributed across sets: Ten of them did test set 1, seven did test set 3 and only five of them test set 3. As a consequence not all results cells are populated equally. However, Kruskal-Wallis test shows that the prominence ratings are independent of the sentence ($p < .27$), therefore we pool the results for the following analysis.

First of all we look at the monotonous audio stimuli. Table 1 shows means and standard deviations of ratings averaged over all subjects and sentences. The outcome suggests a tendency of the inner three syllables to be assigned higher prominence values than the utterance-initial and -final ones.

Table 1: *Listing of means and S.D. of prominence ratings for the monotonous audio-only stimuli (AOM), averaged over the three sentences and all listeners.*

Bens	Haar	war	sehr	lang
Jims	Rad	War	nie	grün
Johns	Bein	War	ganz	blau
1.23/.11	1.43/.06	1.39/.12	1.34/.12	1.16/.07

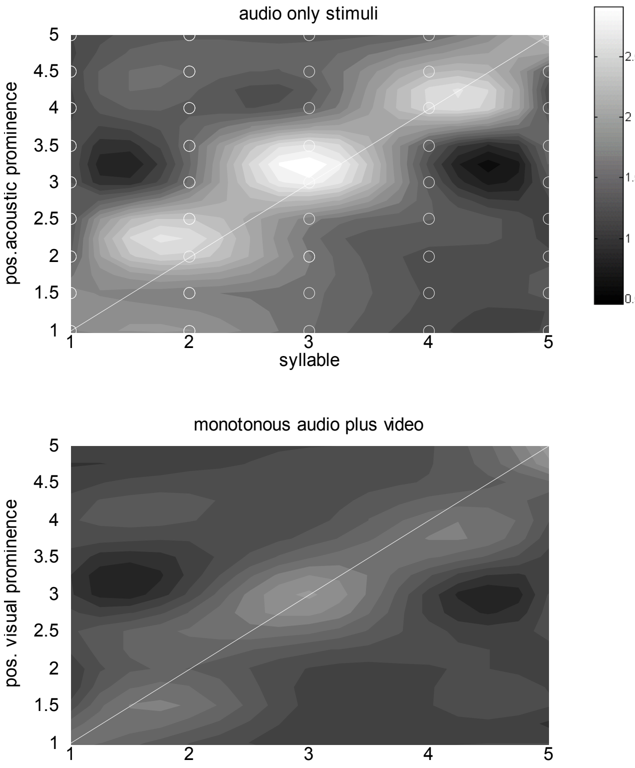

Figure 3: *Visualization of experiment results for stimuli AOF (top) and AVEM (bottom). See text for details.*

Figure 3 displays a graphical representation of prominence ratings for the AOF (top) and AVEM (bottom) stimuli. It was created using *Matlab* employing the *contourf* function, by interpolating over 9x5 matrices of mean prominence results (9 cue positions x 5 syllables) averaged over all subjects and sentences. In the top panel showing the perceived syllabic prominence depending on the position of the *F0* peak, the syllables are aligned along the x axis, and the nine acoustic cue positions along the y axis. Hence, an acoustic position of 1 means that the *F0* peak occurred at the center of syllable 1, acoustic position of 1.5 indicates the *F0* maximum aligned with the boundary between the first and the second syllable. The points of measurement that is, the 9 x 5 matrix, are indicated by white circles. The graphs were created by applying spline interpolation on a mesh grid with a resolution of .25 along both the x axis and the y axis. The result of the interpolation is then mapped onto a 20 level grayscale. Higher prominence values are represented by brighter shading as can be seen from the colourbar at the top right of the figure. In addition, the locations of the acoustic prominences are marked by the diagonal white line. As can be seen, the prominence regions indicated by oval areas of brighter shading move from the left to the right as the acoustic cue position changes. It is also obvious that the prominence region in the center is stronger than towards the left and right edge of the stimulus. Furthermore, the region does not extend symmetrically around the stimulus position indicated by the diagonal, but has a bias towards the left. This suggests that an acoustic cue aligned with the right edge of the syllable has a stronger effect on that syllable than one at the left edge.

Figure 3 (bottom) shows the visualization for monotonous audio combined with eye brow raises (AVEM). In this case the y axis indicates the positions of the visual prominences. Although the regions of prominence develop around the stimulus positions in a similar ways as for the AOF stimuli, it can be seen from the darker shading that the acoustic cue has a much larger impact on perceived prominence than the visual cue. From the alignment of the prominence regions with respect to the visual stimulus a slight bias towards the right can be observed.

Figure 2 shows the same type of visualization for the AVEF stimuli. As in Figure 3 (bottom) the visual prominence is indicated by the diagonal white line. Each of the nine panels represents one of the nine positions of acoustic prominence whose location is also marked by the white vertical line. As expected, the strongest prominence results when both, acoustic and visual stimuli are in the same location. As the visual cue wanders away from the acoustic one, the region of prominence widens and loses intensity, indicated by the darker shading of the peak values. If the visual cue is located far enough from the acoustic one, it develops a region of prominence of its own (see for instance the right-most bottom panel where the acoustic prominence is located in the center of the last syllable). Depending on the case there must be at least one syllable between the two positions for this to happen.

When there is less than a full syllable between the acoustic and the visual cue the perceived prominence is shifted from the position of the acoustic cue towards the position of the visual cue as can be seen in the upward opening angle between the white lines in panels 3 and 4 in the downward opening angle in panels 5, 6, 7 and 8. As already stated, the impact of the acoustic cue at the utterance edges is weak compared with other positions (see panels 1, 2 and 9, panel 8 showing strong perceived prominence on syllable 4). In these cases, the visual cue seems to have a stronger effect.

Table 2: *Listing of means, standard deviation of prominence ratings and N of syllable tokens for the AVEF stimuli depending on the alignment of the acoustic and visual cue.*

acoustic cue	visual cue	mean	s.d.	N
none	none	1.21	.19	618
	on right boundary	**1.40**	.32	75
	on left boundary	**1.57**	.32	75
	in center	1.68	.35	96
on right boundary	none	1.98	.37	75
	on right boundary	**2.16**	.35	12
	on left boundary	**2.39**	.27	9
	in center	2.28	.27	12
on left boundary	none	1.42	.24	75
	on right boundary	1.71	.35	9
	on left boundary	1.63	.46	12
	in center	1.75	.30	12
in center	none	2.08	.39	96
	on right boundary	**2.30**	.40	12
	on left boundary	**2.44**	.27	12
	in center	2.48	.27	15

Table 3: *Listing of means, standard deviations of prominence ratings and N of syllable tokens for the AVNF stimuli (top) in comparison to the corresponding AVEF stimuli (bottom).*

acoustic cue	visual cue nod	mean	s.d.	N
none	none	1.19	.18	77
	on right boundary	**1.42**	.27	10
	on left boundary	**1.36**	.28	10
	in center	1.79	.17	11
in center	none	2.18	.36	18
	on right boundary	**2.60**	.23	3
	on left boundary	**2.39**	.19	3
	in center	2.50	.27	3

acoustic cue	visual cue eye brow	mean	s.d.	N
none	none	1.20	.20	234
	on right boundary	**1.33**	.36	27
	on left boundary	**1.49**	.35	27
	in center	1.58	.34	36
in center	none	2.26	.32	54
	on right boundary	**2.37**	.32	9
	on left boundary	**2.48**	.27	9
	in center	2.59	.21	9

However, there also seems to be a considerably amount of noise in the data, as areas of increased prominence appear in unexpected areas, for instance, in the region in the right bottom corner of the center panel where the acoustic stimulus occurs in the center of the third syllable and the visual cue at the beginning of the utterance.

We subsequently examined the means and standard deviations of perceived prominence for the different alignment situations. For this analysis, we first averaged the syllable-based ratings over all subjects and then averaged over all types of syllables. Table 2 shows the results. When both acoustic and visual stimuli occur in the center of the syllable, the highest prominence ratings are achieved. For the acoustic cue, alignment with the right syllable boundary yields the next-best results.

In contrast, for the visual stimulus, alignment with the left boundary seems to yield more prominence (figures in the table marked in bold), except for the, however, somewhat dispreferred case that the acoustic landmark sits on the left boundary of the syllable (Mann-Whitney U-test of independent samples yields $p < .031$).

Turning to the stimuli using nod movements, we compare them with the corresponding stimuli exhibiting eye brow movements. As mentioned before, the nod stimuli were only created for the acoustic positions in the centers of the 2nd, 3rd and 4th syllable in one of the sentences. Table 3 lists means, standard deviations and N of syllable tokens for the AVNF stimuli (top) and for the corresponding AVEF stimuli (bottom). Despite the larger displacement and duration of the nod movement, the added prominence is in the same range as that for the eye brows. However, the preference for alignment with the left boundary which our results suggested for the latter, does not seem to be replicated for the nods (relevant figures highlighted in bold). Here the right boundary seems to be preferred. Due to the small number of instances, however, this result must be treated with caution.

To round off the analysis and determine the relative contributions of the factors (1) position of acoustic cue with respect to the syllable, (2) position of visual cue, (3) position of syllable, to the prominence rating, we ran an ANOVA the results of which are presented below:

Prominence Rating	variance explained	df	F	Sig.
pos. acoustic cue	51.3%	3	425.51	.000
pos. visual cue	12.8%	3	59.20	.000
syllable position	10.2%	4	34.18	.000

As expected, the position of the acoustic prominence explains most of the variance, followed by the position of the visual one and the particular syllable. The latter result is probably due to the syllables on the stimulus edges receiving lower prominence than the three central ones.

4. Discussion and Conclusions

This study concerned the perception of prominence in auditory-visual speech perception. We constructed five-syllable A/V stimuli in which every syllable in the sentence was a candidate for receiving stress. In various combinations of A/V content subjects had to rate the prominence of syllables on a scale from 1-3. Results show that the acoustic prominence outweighs the visual one, and that the integration of both in a single syllable is the strongest when the *F0* peak or the point of maximum displacement, respectively, are aligned with the center of the syllable. However, *F0* peaks aligned with the right boundary of the accented syllable and, in contrast, the maximum of the eyebrow displacement aligned with the left boundary also boost prominence considerably. Nods seemed to cause a similar amount of additional prominence as eye brow raises, despite their longer duration and stronger influence on the optic flow. There seems to be a preference for the nods to be aligned with *right* boundary. This perceptual difference compared to the eye brow movements seems to be confirmed by recent auditory-visual production results by Kim et al. [11] on natural speech who found that the amount of eyebrow movement in narrow focus condition tended to be larger before or at the focused item than after it. Head rotation (nodding) tended to occur later, being maximal in the critical region and still occurring often in the post-critical one.

It must be stated that the experiment task was a rather taxing one since each of the syllables had to be rated. Choosing the two most prominent syllables might have yielded more consistent results. Furthermore, the artificiality and uniformity of the stimuli is likely to have cause repetitious replies in some of the subjects. Some of them commented that they went with the acoustic stimulus most of the time and rarely took notice of the visual cue.

In future work we will compare the results of the native German subjects with those of the native Turkish ones which represent the second largest group of our participants and whose data we so far did not evaluate, as well as with other language groups in order to see whether alignment preferences are universal or culturally dependent. It will also be important to investigate in further detail the differences between the perception of eye brow and nod movements. To this end we will construct datasets which are better balanced than the current one. Furthermore we are interested in the relationship between the magnitude of eye brow displacement/height of the *F0* peak and the degree of perceived prominence.

5. Acknowledgements

Thanks go to Philipp Hempel who programmed the experiment software and all the students at Beuth University who took part in the experiment. I would also like to thank Marijo Mateljan for providing the “face” for the stimuli.

6. References

[1] McGurk, H., & MacDonald, J. (1976). Hearing Lips and seeing voices. In: *Nature*, Band 264, S. 746-748.

[2] Mixdorff, H., Charnvivit, P. and Burnham, D. (2005): Auditory-Visual Perception of Syllabic Tones in Thai. In Proceedings of AVSP 2005, pp. 3 - 8, Parksville, Canada.

[3] Swerts, M. & Krahmer, E. (2008). Facial expressions and prosodic prominence: Comparing modalities and facial areas. Journal of Phonetics, 36(2), 219-238.

[4] Hönemann, A. & Mixdorff, H. & Fagel, S. (in press). A preliminary analysis of prosodic features for a predictive model of facial movements in speech visualization. Proceedings of Nordic Prosody 2012, Tartu, Estonia.

[5] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., van der Vreken, O. (1996). The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. Proc. ICSLP, Philadelphia, USA, 1393-1396.

[6] Boersma, P., & Weenink, D. (2012). *Praat: doing , phonetics by computer* (Version 5.1.26) [Computer program]. Retrieved April 4, 2012, from <http://www.praat.org/>.

[7] Fujisaki, H. and Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. Journal of the Acoustical Society of Japan (E) 5(4): 233-241.

[8] Mixdorff H. (2000). A novel approach to the fully automatic extraction of Fujisaki model parameters. Proceedings of ICASSP 2000, vol. 3, 1281-1284, Istanbul Turkey.

[9] Mixdorff, H., “FujiParaEditor”, http://public.beuth-hochschule. de/~mixdorff/thesis/fujisaki.html, 2009.

[10] VirtualDub version 1.8.8, http://www.virtualdub.org, retrieved 2.2.2009.

[11] Kim, Cvejic, E. & Davis, C. (in press). Tracking eyebrows and head gestures associated with spoken prosody. Submitted for: Speech Communication Special Issue on Gesture and Speech in Interaction.