ALIGNMENT BETWEEN RIGID HEAD MOVEMENTS AND PROSODIC LANDMARKS

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

Beuth University of Applied Science Berlin[1], Zoobe message entertainment GmbH[2]
{ahoenemann|mixdorff}@beuth-hochschule.de, fagel@zoobe.com

Abstract: In our study we recorded and analyzed an audiovisual speech corpus to develop a model which predicts head and facial non-verbal movements accompanying speech. The model is intended to improve the naturalness of avatars. Our previous paper already gives a preliminary analysis of our speech corpus which includes acoustic and visual recordings of seven individual speakers who talk about three minutes about their last vacation. We showed that for each speaker 20-30% of events in each motion class are aligned with prominent syllables in phrase-initial or -medial position and that the speakers moved most often at the end of an intonation phrase. We also observe that the speakers differ in strength and frequency of visible events. However, there is also a great ratio of about 60% of motion events which are not assigned to the target syllables. In order to account for this result, further analyses had to be carried out. The present paper shows further analyses of the relationship between speech and movements. Therefore, we extracted the fundamental frequency (F0) and the intensity of the acoustic signals using Praat. By marking the prominent syllables we obtained a description of the course of F0. We use the Principle Component Analysis (PCA) to determine the linear combinations of the visual parameters that constitute the main head movements.

1. Introduction

Besides the linguistic content the facial and head movements of speakers provide a lot of information for a better understanding of what s/he wants to say. Several investigations confirm this. For example Al Moubayed, Beskow and Granström showed that the visual cues are reinforced by the perception of prominences. Their experiment presented acoustic speech signals with a talking head and found that the test subjects perceived the accented word better when head and eyebrow movements were synchronized with prominence markers in the acoustic speech signal. In contrast it was difficult to hear the audible accents by a talking head with a neutral facial expression [1]. This indicates that speech is a multi-modal process.

A realistic simulation of a talking head is still a challenge and therefore subject of many investigations. The aim of this study concerns the modeling of prosodic features to predict visual cues aligned with the acoustic signal. On that account we need a good understanding about the audiovisual relationships. Our previous results showed that many non-verbal movements are due to idiosyncrasy [3]. Further studies are presented in the present paper. A detailed investigation of the alignment between prosodic features such as the fundamental frequency and the intensity of speech and the motions which we observed yields a better understanding of this relationship. An objective analysis of the motion capture data supplied the visual parameters based on Principle Component Analysis (PCA).

Our acoustic analysis includes standard features such as the F0 range, F0 maximum and F0 minimum, as well as the investigation of the intensity range, maximum and minimum. The syllable duration as an indicator of prominence was also of interest. In addition to the prominent syllable proper considering the syllables before and after was regarded as important. The segmentation of F0 and intensity contours according to accented and unaccented ranges provides information about the interrelation between prosodic features and movements.

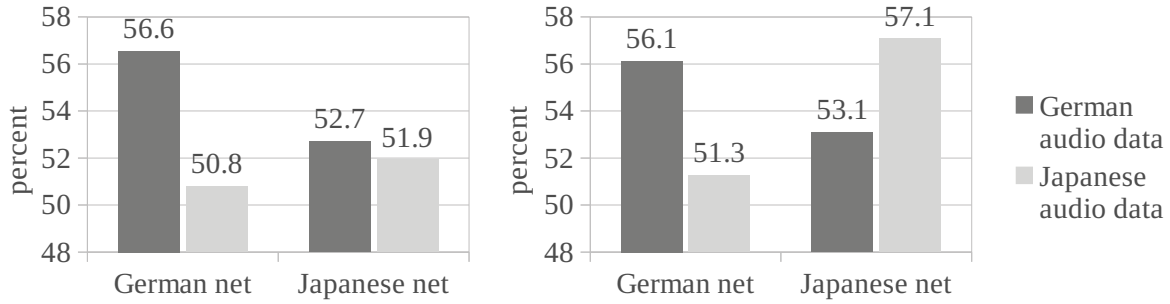

Figure 6: Results for ESNs applied to data of the same and of the other culture, on the left hand for "na", on the right hand for natural feedback.

5 Discussion

Our approach worked for "na" utterances as well as for utterances in natural language. From male persons the natural feedback could be recognized even better, while from females the "na" condition was easier to classify. Although not significant, these differences fit the experimenter's observations that most female persons easily used prosody in "na" feedback, while some males had difficulties to convey their intent without using words.

Table 1: Classification of natural data with "na" trained networks: Percentages of subjects that reach an accuracy higher than 50 / 60 / 70 / 80 %.

Mean performance	67%
Standard deviation	13.9
better than 50%	92%
better than 60%	67%
better than 70%	46%
better than 80%	21%

While some test subjects produced speech including high prosody, others relied more on lexical information using adult-directed speech. This could be prevented in further studies by developing the experimental design more carefully. For Flobi it has not been proved that he triggers infant-directed speech, so he might have been perceived by the participants in different ways. It might also be an idea for improvement to choose a more child-like voice for the robot's utterances.

The usage of prosody in feedback appears to be also highly situation dependent. Concerning the scenario, a task without any usage of lexical information might be better suited to enforce more emotional responses. Also a continuous task where the user's response directly influences the robot's next action could yield stronger prosody.

As the generalization properties seem to depend on the quality of the training data, training with data from actual infant-directed speech might be useful and could make it possible to also classify cases with more subtle prosody.

MFCCs are the standard approach for speech recognition, but are also often included additionally to the prosodic features for emotion recognition tasks. That MFCCs alone proved to work so well might be, because the MFCCs contain the whole speech signal, which can not be reconstructed through the prosodic features alone.

Looking back at our research questions we can conclude that differences between German and Japanese users seem to exist. A network trained on German data works significantly better on German audio data than on Japanese audio data, but a significant difference cannot be found the other way round. This could result from the fact that the prosody in Japanese was often very indistinct, at least in the explored group. The reason for that might be cultural influences, as a clear "No" is rarely used in Japanese everyday language.

Furthermore, it is possible to recognize positive and negative feedback from utterances in a speaker-dependent way. To some extent also speaker-independent networks would be possible to implement; but we observe very diverse strategies of giving feedback intra-culturally as well as inter-culturally. This indicated that a general solution is difficult to achieve and a speaker-dependent model for classification would be the better choice.

References

[1] BOTINIS, A., B. GRANSTRÖM and B. MÖBIUS: *Developments and paradigms in intonation research*. Speech Communication, 33(4):263–296, 2001.

[2] BREAZEAL, C. and L. ARYANANDA: *Recognition of affective communicative intent in robot-directed speech*. Autonomous robots, 12(1):83–104, 2002.

[3] DAI, J., G. VENAYAGAMOORTHY and R. HARLEY: *An introduction to the echo state network and its applications in power system*. In *ISAP'09. 15th International Conference on Intelligent System Applications to Power Systems, 2009*, pp. 1–7. IEEE, 2009.

[4] FERNALD, A.: *Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages*. Child development, 64(3):657–674, 1993.

[5] FERNALD, A., T. TAESCHNER, J. DUNN, M. PAPOUSEK, B. DE BOYSSON-BARDIES, I. FUKUI et al.: *A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants*. Journal of child language, 16(3):477–501, 1989.

[6] FRIEND, M.: *The transition from affective to linguistic meaning*. First Language, 21(63):219–243, 2001.

[7] FRIEND, M. and J. BRYANT: *A developmental lexical bias in the interpretation of discrepant messages*. Merrill-Palmer Quarterly: Journal of Developmental Psychology, 2000.

[8] HOLZMANN, G.: *Echo State Networks with Filter Neurons and a Delay&Sum Readout with Applications in Audio Signal Processing*. Master's thesis, Graz University of Technology, Austria, 2008. http://aureservoir.sourceforge.net.

[9] JAEGER, H.: *The "echo state" approach to analysing and training recurrent neural networks*. GMD Report 148, GMD - German National Research Institute for Computer Science, 2001.

[10] LUETKEBOHLE, I., F. HEGEL, S. SCHULZ, M. HACKEL, B. WREDE, S. WACHSMUTH and G. SAGERER: *The bielefeld anthropomorphic robot head "Flobi"*. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 3384–3391. IEEE, 2010.

[11] MORTON, J. and S. TREHUB: *Children's understanding of emotion in speech*. Child development, 72(3):834–843, 2003.

[12] SKOWRONSKI, M. and J. HARRIS: *Automatic speech recognition using a predictive echo state network classifier*. Neural networks, 20(3):414–423, 2007.

[13] THOMSON, D. and R. CHENGALVARAYAN: *Use of periodicity and jitter as speech recognition features*. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998*, vol. 1, pp. 21–24. IEEE, 1998.

This paper is structured as follow: Section 2 describes the experiment setup. In Section 3 we outline the classification of the syllables and the segmentation of F0 and intensity contours, as well as the results of the statistical analysis of the acoustic and visual data. Section 4 contains the results of the audiovisual analysis. The calculated correlations between the acoustic and visual data are presented. Section 5 provides the conclusions of our paper.

2. Experiment Setup

For our experiment we examined three male and four female adult native speakers of German. We used a dataset of a free narrative where the speakers talked about their last holiday. The free narrative has the advantage that the spoken passage provides a great range of prosodic features. They contain sentences of different lengths, breaks, hesitations. In addition, speakers showed a wide variation of movements.

To capture the motion data of the speakers we used the Qualisys motion capture system [5]. Three infra-red cameras scanned the passive markers which were attached on the face and the head of the speakers with a frame rate of 60 frames per second. These data are processed by the software Qualisys Track Manager (QTM), which presents them in a 3D coordinate system. Synchronously we recorded the acoustic speech signals at 44.1kHz, 16-bit mono with an externally connected microphone along with Mini-DV video. For the analysis we resampled the audio stream to 16kHz.

43 markers appeared sufficient to capture all the important motions of the face and head. We placed three markers with a diameter of 10mm for the rigid movements, which were attached to a head rig, and 40 markers for the non-rigid motions. These markers had different sizes depending on the facial region to be measured. For the fine lip movements we used markers of size 2.5mm, and for the other moving regions like the chin, nose, cheeks and the eyebrows we used marker of size 4mm. For the regions with less movement, like the forehead, temples, bridge of the nose and the throat a marker of average a size of 7mm was appropriate. Figure 1 shows the marker configuration on two of our test subjects**.**

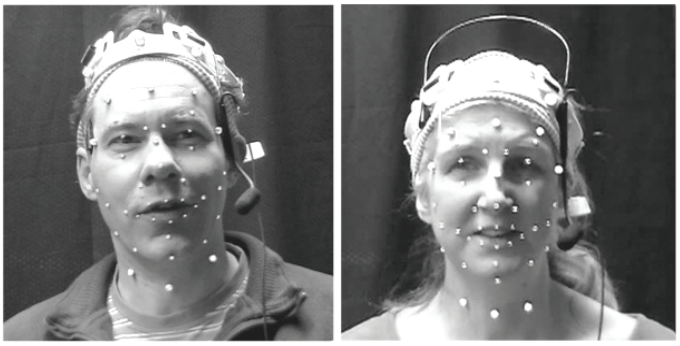

Figure 1: Marker placed on the face and head of two of our speakers

3. Data Preparation and Material

3.1 Acoustic Data

The acoustic data was annotated at the word and syllable level. We segmented about one minute of the recording of each speaker and perceptually determined the prominent syllables and phrase breaks using Praat [2]. Our interest was to consider only the speech syllables, therefore breaks and hesitations were eliminated from the text; they will be investigated separately.

Our empirical observation was that our speakers' movements often started or ended prior to or following an accented syllable. For this reason we also have to consider the syllables before and after accented syllables. On the one hand we investigated the syllables with respect to the

superordinate phrase and on the other hand we investigated the syllables are treated independently of phrases. Table 1 shows this classification and the meaning of them.

Syllable classes phrase dependently	
A	Accented syllable, phrase-initial or medial position
B	Unaccented syllable phrase-finally
A/B	Accented syllable phrase-finally

Syllable classes phrase independently	
ACC	Accented syllable
UNACC	Unaccented syllable
PRE	Unaccented syllable before an accented syllable
POST	Unaccented syllable after an accented syllable
PREPO	Unaccented syllable between two accented syllable

Table 1: Syllable classes and their description

Our analysis of the syllable duration confirms findings from other studies that indicate increased durations when a speaker emphasizes a syllable or when it is the last syllable of a phrase. As can be seen in Figure 2 on the left side, the average duration of all speakers was longer at A (211ms, s.d. 35ms), B (232ms, s.d. 407ms) and A/B (366ms, s.d. 655ms) syllables as at an unaccented syllables (188ms, s.d. 253ms). Conspicuously, emphasized syllables phrase-finally are significantly longer than all other syllables. The right side of Figure 2 shows the average syllable durations of all speakers and as expected the duration was longer if the syllable was adjacent to an accented syllable, that is, the PRE (156ms, s.d 186ms), POST (172ms, s.d. 181ms) and PREPO (167ms, s.d. 126ms) syllables. The accented syllables had the longest duration (252ms, s.d. 42ms), however, the duration of an unaccented syllable (140ms, s.d. 28ms) was much shorter.

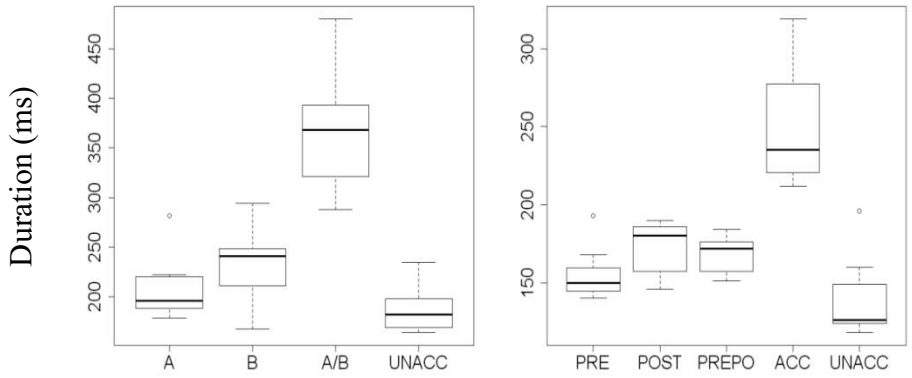

Figure 2: Duration in ms of different syllable classes of each speaker as describe in Table 1

We extracted the fundamental frequency (F0) and the intensity of the speech signals with a time step of 16 ms to match the visual data time step. We chose a pitch range for males between 50 – 350Hz and for females between 75 - 400Hz. After manually corrected errors such as octave jumps or erroneous measures due to creaky voice we interpolated the F0 to calculate the unvoiced sections and transformed it into z-score.

We divided the F0 and the intensity curve into two different sequence classes. The first sequence, labeled as AccSeq, is the sequence from a PRE or ACC syllable to the next PRE syllable and includes at least one accented syllable. The other sequence, which includes only unaccented syllables, is the sequence from a UNACC syllable to the next PRE or ACC syllable; we labeled it as UnAccSeq. Due to the removal of the hesitations and breaks, not all syllable classes are necessarily included in the sequences e.g. it could be that there are no PRE or POST syllables.

We computed the duration of the syllables included in each sequence. As expected the duration of the syllables included in an AccSeq (184ms, s.d. 49ms) were on average longer than the duration compared to syllables included at an UnAccSeq (124ms, s.d. 47ms). Furthermore, we identified the maximum and minimum of the F0 and the intensity contour of each sequence class to calculate the F0 and intensity range. The difference of the maximum and minimum gives the frequency of the range classes.

The average value of all speakers at F0 was at an $AccSeq_{F0}$ (69.547Hz, s.d. 52.161Hz) much higher than at an $UnAccSeq_{F0}$ (12.936Hz, s.d. 9.726Hz), the intensity range shows the same tendency, the $AccSeq_{Int}$ (29.017Hz, s.d 6.725Hz) exhibited higher frequency as at an $UnAccSeq_{Int}$ (17.439Hz, s.d. 9.655Hz). The standard deviation of the intensities shows that there is slightly more variation at an unaccented range. Figure 3 shows the F0 and intensity range of each speaker at an AccSeq and at an UnAccSeq. In general there is a greater variation in F0 at sequences which include accented syllables, however, unaccented sequences were spoken more monotonously.

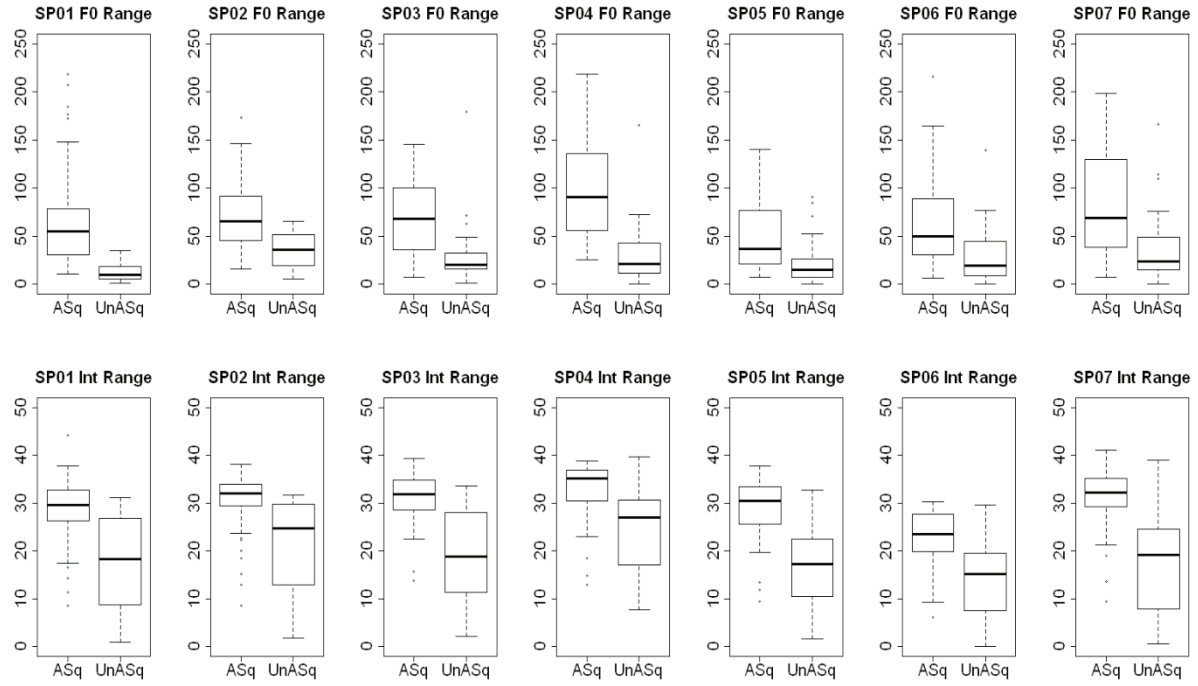

Figure 3: F0 range and intensity range in Hz at AccSeq (ASq) and UnAccSeq (UnASq) classes of each speaker

The following examples show that there are strong correlations of F0 and intensity features such as the maximum and minimum in relation to the syllable duration calculated for each sequence: SP01: $F0_{MAX}Syl_{DUR}$: .625**, $F0_{RANGE}$:$Syl_{DUR:}$ .661**, $Int_{MAX}Syl_{DUR:}$ .559**, SP03: $F0_{MIN}Syl_{DUR}$: -.321**, $Int_{MIN}Syl_{DUR}$: -.534**, SP07: $F0_{RANGE}Syl_{DUR}$: .499**, $Int_{MIN}Syl_{DUR}$: -.476**, $Int_{RANGE}Syl_{DUR}$: .593** ($p < 0.01$ **).

3.2 Visual Data

The movements of the speakers were classified and annotated on the digital video in Anvil [4]. We defined different motion classes for the main facial regions e.g. the lips ((L)-Down) and eyebrows ((EB)- Raise), for the head motions we defined e.g. (H)- BackUp, (H)-SideTurn and (H)- Forwards. The head movements are our main interest in the present paper.

We grouped four main types of head movements depending on their orientation. Table 2 lists the classes and the descriptions of them. Table 3 gives an overview of the proportion of these motion classes for each speaker. We also computed the average duration of the motion classes

of each speaker and found that LRR motions (1188ms) on average have the longest and FBS motions (555ms) the shortest duration. Table 3 shows this motion duration and the standard deviation on average for each speaker.

Motion classes of main head movements	
UDT	up and down turn e.g (H)- BackUp, (H)- Down, (H)- Nod
LRT	left and right turn e.g (H)-SideTurn, (H)- SideTurn-R
FBS	forwards and backwards shift e.g. (H)- Forwards, (H)- Backwards
LRR	left and right rock e.g. (H)- SideRock, (H)- SideRock-R

Table 2: Motion classes of the main head movements and their description

Speaker/Motion		*UDT*	*LRT*	*FBS*	*LRR*
SP01	*%*	53.1	6.2	25.0	15.6
	mean/s.d. [ms]	816/608	700/424	255/120	1188/980
SP02	*%*	20.0	21.8	47.3	10.9
	mean/s.d. [ms]	620/230	559/314	446/250	584/245
SP03	*%*	48.7	35.9	15.4	---
	mean/s.d. [ms]	1086/518	1007/625	714/363	---
SP04	*%*	45.2	22.6	25.8	6.5
	mean/s.d. [ms]	852/260	806/300	813/477	519/226
SP05	*%*	34.1	4.9	34.1	26.8
	mean/s.d. [ms]	963/923	543/250	556/310	490/203
SP07	*%*	33.3	29.4	13.7	23.5
	mean/s.d. [ms]	738/593	1071/1418	473/156	828/1012
SP07	*%*	38.3	33.3	25.0	3.3
	mean/s.d. [ms]	1054/816	861/713	624/262	880/--

Table 3: Percentage ratio of total motion events of each motion class in relation to head movements, Mean Duration and standard deviation in ms of each speaker and motion class

For a closer analysis we only used the underlying motion capture data of the visually detectable motions. On that account we defined a rigid body from three markers on the forehead. We derive six degrees of freedom, three for rotational movement and three for translational movements. The pitch angle describes the rotation around the x- axis, the y-axis is specified by the yaw angle and the roll angle describes the rotation around the z-axis. The translational movements are described by x, y and z.

We used the Principle Component Analysis (PCA) to determine the linear combination of the motions of each speaker. The inputs for the PCA were the three rotational and three translational parameters. The results of PCA for each speaker are shown in Figure 4. Three PCs explain between 86% and 95% of the variance of speaker movements.

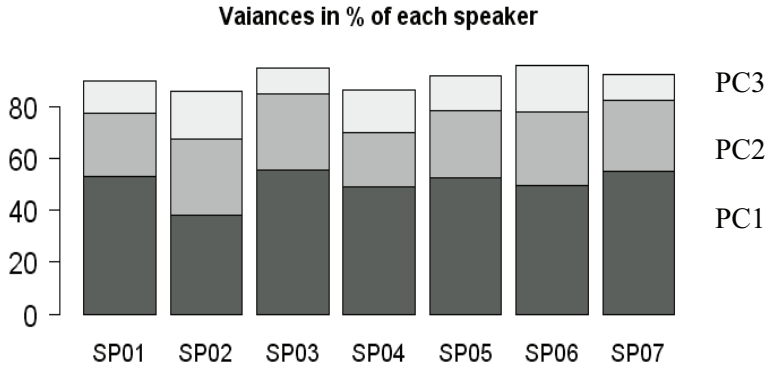

Figure 4: F0 Variance in % from each of the speaker and PCs

The computed components of the PCA are in general difficult to interpret, that means in our case that the components do not relate directly to any specific class of motion. Therefore a comparison of the components with our visually detected and labeled motion events was helpful. Further information was given by the correlations between the three components and the 6DOF original input data. The strongest correlation with the first component was shown in relation to x translational movements and the roll angle (rotation around the z-axis), however, the second component had a strong correlation with z translational movements and the yaw angle (rotation around the y-axis), the third component with y translational movements and the pitch angle (rotation around the x-axis). This leads to the assumption that PC1 mostly includes information of LRR/FBS motions, PC2 information of LRT motions and PC3 information of UDT motions.

4. Results

45% of all motion events begin at an ACC syllable (accented syllable) and to 22.5% at PRE syllables (syllable before an accented syllable), however, only 16.0% start at a POST syllable (syllable after an accented syllable) and at an unaccented syllables 10.1%. POST account 29.9% of movement offsets and unaccented syllables 30.9%. 37.5% of movements terminate at an ACC syllable.

Phrase offsets often coincide with movement offsets. 20.1% of these accord at A/B syllables (accented syllable phrase-finally) and 24.3% at B syllable (unaccented syllable phrase-finally). However, speakers started seldom a movement at the end of a phrase. 11.2% at A/B syllables and only 3.6% at B syllables.

We calculated the correlation between the audiovisual features for instance the F0 over the whole dataset of each speaker as follows: $SP01_{F0PC1}$: -.257**, $SP01_{F0PC2}$: .230**, $SP02_{F0PC2}$: .343**, $SP04_{F0PC1}$: .218**, $SP05_{F0PC2}$: .226**, $SP07_{F0PC2}$: .268** ($p < 0.01$ **).

In addition we calculated the correlation of F0 and intensity in relation to the computed PCs only at the perceived motion events UDT, FBS, LRT and LRR of each speaker. The correlations are listed in Table 4.

Speaker	*P*	*UDT*		*FBS*		*LRT*		*LRR*	
/Motion	*C*	*F0*	*INT*	*F0*	*INT*	*F0*	*INT*	*F0*	*INT*
SP01	*1*	-.320**		-.424**				-.517**	
	2	.235**		.324**		.571**			
	3					-.555**	-.226*		
SP02	*2*	.524**	.205**	.424**		.202**		.339**	.200**
SP03	*1*			.385**					
	2			.308**					
SP04	*1*	.214**		.365**				.771**	
	2	.310**		.352**					
SP05	*1*				.229**		.388**		
	2			.251**		.362*	.485**	.289**	
	3			-.222**	-.236**		.466**		
SP06	*1*								-.344**
	2		.284**		.-288**				
	3			-.323**					
SP07	1	.280**		.536**		.204**			.330*
	2					.527**	.201**		-.295*
	3			.222**					

Table 4: Correlations between the F0 values at the detected motion events and the principle components of each speaker ($p < 0.01$ **, $p < 0.05$ *)

We also estimated the correlation between the F0 and intensity features such as the maximum, minimum and range in relation to features of the PCs. To this end we did the same analysis with the PCs as with the acoustic data. We identified the maximum and the minimum of the PCs for each of AccSeq (at least one accented syllable) and UnAccSeq (without accented syllables) and computed the ranges between them. Table 5 shows the correlation of the comparison. The correlations were computed of the total dataset of our seven speakers.

	F0 max	*F0 min*	*F0 range*	*Int max*	*Int min*	*Int range*	*Syl. dur*
PC1 max							
PC1 min							-.246**
PC1 range			.343**		-.330**	.376**	.449**
PC2 max	.212**		.268**	.209**			.252**
PC2 min							-.267**
PC2 range			.325**	.327**	-.230**	.356**	.527**
PC3 max							.308**
PC3 min					.206**	-.236**	-.285**
PC3 range			.325**	.259**	-.323**	.402**	.569**

Table 5: Correlations between the acoustic and visual feature of all seven speakers ($p < 0.01$ **)

The ranges of the PCs could be seen as the activity strength of the speakers. The average value of each PC of all speaker at the different ranges are follow: $PC1_{ACC}$: 1.472 (s.d. 1.270), $PC1_{UNACC}$: 0.485 (s.d. 0.594), $PC2_{ACC}$: 0.878 (s.d. 0.624), $PC2_{UNACC}$: 0.335 (s.d. 0.387), $PC3_{ACC}$: 1.470 (s.d. 1.278), $PC3_{UNACC}$: 0.524 (s.d. 0.563).

Figure 5 shows as an example the calculated range of the first component of each speaker at an accented and unaccented sequence. Clearly the speaker show more activity at an AccSeq. The same tendency is apparent for the second and third component.

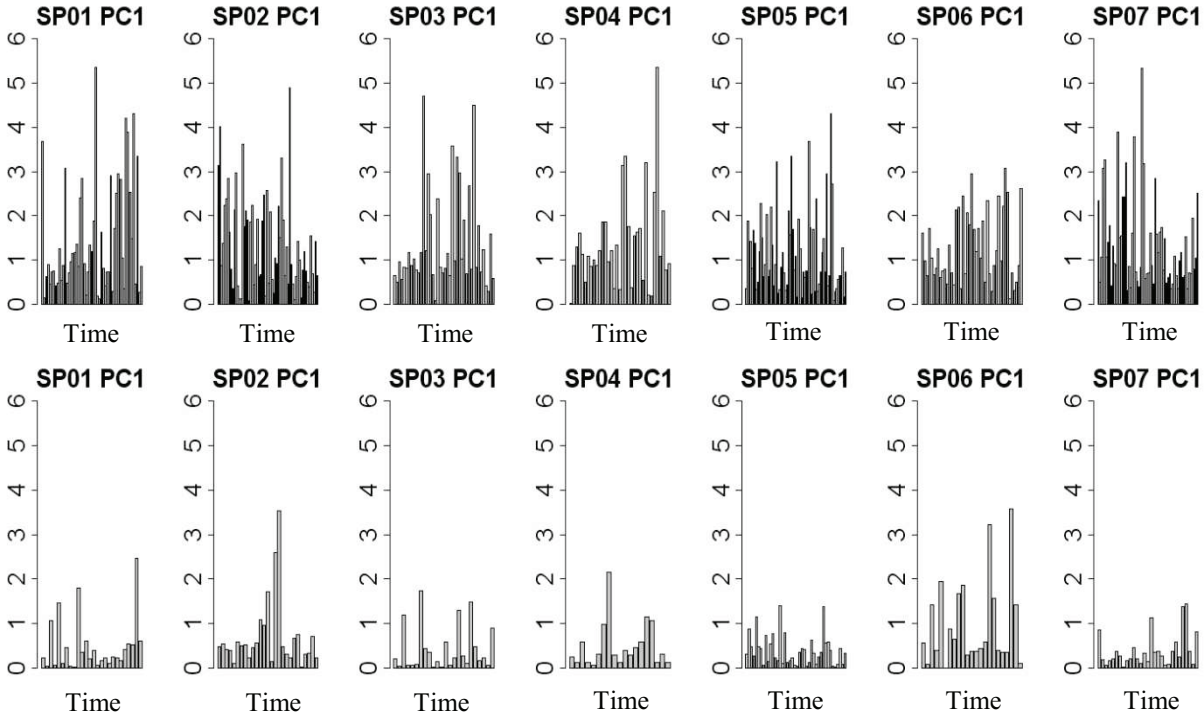

Figure 5: PC1 range as activity strength of each speaker at AccSeq (above) and UnAccSeq (below)

5. Conclusion

The aim of the analysis of prosodic features, such as the fundamental frequency and intensity as well as the syllable duration, was to find relevant parameters for a predictive model whose outputs are appropriate visual cues.

It has been pointed out that syllable duration correlates strong with the visual features of our principle component analysis such as the ranges, maxima and minima of the PCs. Clearly, syllable duration is an indicator of prominence because the results show that syllable duration is longer at an accented syllable than at unaccented syllables. Not only is prominence crucial for this delay but also the position of a syllable e.g. a syllable before or after an accented syllable exhibits longer duration. It has been shown that the speaker up to 22% of their motion started at an unaccented syllable before an accented syllable (PRE) and up to 29% their movements ends at an unaccented syllable after an unaccented syllable (POST). Therefore, these findings are important because the compliance only of emphasis is not sufficient.

The segmentation of the F0 and the intensity into accented and unaccented sequences was helpful to examine the influence of the prominences and the nearest environment on visual behavior. As it turned out, at unaccented sequences the speaker showed significantly fewer activities than at accented sequences. This result indicates that the curve of the F0 has a great influence on the degree of activity of the speaker.

The calculated correlations show that there exists an alignment between the prosodic features such as the maximum, minimum, range and the main head movements. The visualization of motion through the modeling of prosody of speech is a challenge. Our results supply a great basis for the realization but there are many influencing factors which have to be examined in further studies.

Acknowledgment

The first author is funded by the European Social Fund (ESF) and supported by the Berlin Senate for Economics, Technology and Research.

Bibliography

[1] Al Moubayed, S., Beskow, J., Granström, B..Auditory visual prominence. From intelligibility to behavior. J Multimodal User Interfaces, 3: 299–309; 2010.

[2] Boersma, P. & Weenink, D. Praat: doing phonetics by computer (Version 5.3) [Computer program]. Retrieved February 21, 2012, from www.praat.org.

[3] Hönemann, A., Mixdorff, H., Fagel, S. (forthcoming). A preliminary analysis of prosodic features for a predictive model of facial movements in speech visualization. In: Proceedings of Nordic Prosody XI. Peter Lang.

[4] Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multimodal Dialogue. In: *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*.Aalborg, Denmark. pp. 1367-1370.

[5] Qualisys Company Homepage. Retrieved February 10, 2012 from www.qualisys.com