Perception of Prominence Intensity in audio-visual Speech

Manfred Nusseck[1], *Douglas W. Cunningham*[2], *Jan Peter de Ruiter*[3], *Heinrich H. Bülthoff*[1]

[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2]University of Tüebingen, Germany
[3]Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

manfred.nusseck@tuebingen.mpg.de, douglas.cunningham@gris.uni-tuebingen.de
janpeter.deruiter@mpi.nl, heinrich.buelthoff@tuebingen.mpg.de

Abstract

Multimodal prosody carries a wide variety of information Here, we investigated the roles of visual and the auditory information in the *production* and *perception* of different emphasis intensities. In a series of video recordings, the intensity, location, and syntactic category of the emphasized word were varied. Physical analyses demonstrated that each speaker produced different emphasis intensities, with a high degree of individual variation in information distribution. In the first psychophysical experiment, observers easily distinguished between the different intensities. Interestingly, the pattern of perceived intensity was remarkably similar across speakers, despite the individual variations in the use of different visual and acoustic modalities. The second experiment presented the recordings visually, acoustically, and audiovisually. Overall, while the audio only condition was very similar to the audiovisual condition, there was a clear influence of visual information. Weak visual information lead to a weaker audiovisual intensity, while stong visual information enhanced audiovisual intensity.

Index Terms: multimodal speech, audiovisual prosody, prominence

1. Introduction

The multimodal nature of speech prosody is particularly noticeable in emphatic speech. There are several acoustic signals for prominence, including changes in amplitude and changes in the pitch of the fundamental frequency (f0) [1], although the importance of this latter signal is somewhat controversial (see, for example, [2]). Visually, Renwick et al. [3] have shown, for instance, that head and eyebrow motions sometimes indicate prominence. It is also clear that the separate signals in the two modalities are related to each other. Changes in the f0, for example, correlate up to 70% with certain facial (e.g., eyebrow) or head motions [4, 5, 6, 7]. Furthermore, using multimodal signals carries several general advantages for comprehension: Munhall et al. [8] found that the audiovisual presentation of a speaker increased the intelligibility of their sentences over mere auditory presentation. Thus, speech in general, and emphasis in particular, is **produced** multimodally, and this seems to enhance at least some aspects of speech **comprehension**.

Prosody can be used to specify a number of different things. While most previous research has focused on how prosody can determine the focus of a sentence or how it affects sentence parsing, there is some suggestion that there are production differences in emphasis *intensity*. For example, Gussenhoven et al. [9] have shown that speakers vary the height of associated f0 to express different degrees of emphasis. Furthermore, there are some difference in the perception of emphasis intensity: Swerts [10] found that the addition of visual information for emphasis increased the perceived prominence of that word compared to an audiovisual presentation without visual emphasis. It is possible, then, that the fine-grained, physically present, multimodal distinctions in emphasis intensity allow a fine-grained perceptual differentiation of emphasis intensity. Here, we investigated the relative contribution of the visual and auditory channels in both the production and the perception of different emphasis intensities.

1.	Diese Fliesen betonen das moderne Design *These tiles emphasis the modern design*
2.	Dieses Dekor betont das groflächige Format *This decor emphasises the extensive format*
3.	Diese Fliesen ergänzen das dekorative Muster *These tiles complement the decorative sample*
4.	Dieses Muster ergnzt das schlichte Dekor *This sample complements the simple decor*
5.	Diese Farben unterstreichen das moderne Format *These colours underline the modern format*
6.	Dieses Format unterstreicht das einfache Design *This format underlines the simple design*
7.	Diese Farben vollenden das dekorative Design *These colours complete the decorative design*
8.	Diese Fliese vollendet das moderne Muster *This tile completes the modern sample*
9.	Diese Muster verbergen das einfache Format *These samples hide the simple format*
10.	Dieses Format verbirgt das groflächige Dekor *This format hides the extensive decor*

Table 1: The stimulus sentences, with a rough English translation.

2. Audiovisual recordings

2.1. Sentences

We used ten different German sentences (see Table 1), each of which had the same syntactic structure: subject, verb, and then object (SVO). The object consisted of an adjective and a noun. Both singular and plural forms of the verb were used. The same adjectives and nouns were used in multiple sentences in order to increase the semantic similarity of the sentences.

The sentences describe a situation of selling tiles. The sentences were designed to ensure a neutral conversational context, in part by the absence of prior emotional involvement or investment with the conversational topic.

The location and syntactic category of the focus was systematically varied for each sentence: The accent bearing word could be the verb, the adjective, or the predicate noun. Additionally, the emphasis intensity was systematically varied; it could be No, Low, or High intensity. Thus, each sentence was recorded seven times; once each for verb high, noun high, adjective high, verb low, noun low, adjective low, and no intensity. Note that *each recording of a sentence had only one focus location*.

2.2. Speakers

The 70 different stimuli were recorded from each of 8 individuals (4 male and 4 female), all of whom spoke German as their native language, had a standard German accent (without specific dialect), had no obvious speech impediments, and little acting experience. Each speaker was told which syllable of the verb, adjective, or noun was to be emphasized. To help ensure consistency in the location and type of emphasis, each sentence was preceded by a context sentence. For instance, a theme/rheme contrast was set up in the target sentence ”These tiles emphasize the *modern* design” by preceding it with the context sentence ”While some tiles emphasize the *antique* design”. The contrast sentences were not to be spoken aloud and were not recorded. The speakers were asked to express a difference between the intensities, but were asked not to force the emphasis or to artificially exaggerate it. Finally, they were asked to not include segmental durations such as prosodic pauses or rhythmical phrases (see [11]). During the recordings, the speakers were free to move in any way they felt normal and appropriate, but were asked to refrain from placing their hands in front of their faces.

Each sentence was recorded three times in a row with a short pause between the repetitions. The best of the three repetitions (e.g., one containing no mistakes, no emotional expressions, no unusual eye gaze behaviour) was chosen for the experiment.

2.3. Recording Equipment

The recordings were made using the Max Planck Institute for Biological Cybernetics’s VideoLab. The setup includes six recording units, each of which consists of a digital video camera, a frame grabber, and a computer (for more details see [12]). The present recordings were made at 50 frames/s using two of the recording units, with the main camera being positioned in front of the speaker. During the recordings the speakers wore a black hat with six green dots on it. These points were used to track the head position and its movements in post-processing analysis.

3. Physical analysis of the recordings

In a first step, we examined the production of emphasis to address two questions. First, were the speakers actually able to produce different levels of intensity? Second, if they did produce different intensity levels, where was the information for intensity level located? For these analyses, we focused on three phonetically-coupled prosody parameters: fundamental frequency, voice amplitude, and rigid head motion. For the remainder of this paper, these three parameters or information channels will be referred to as *semiotic channels*, in accordance to the definition given by [13]. For these three semiotic channels, we determined the peak value for the Low and the High intensity conditions at the prominence point. We took the values at the same points from the sentences with No prosodic focus.

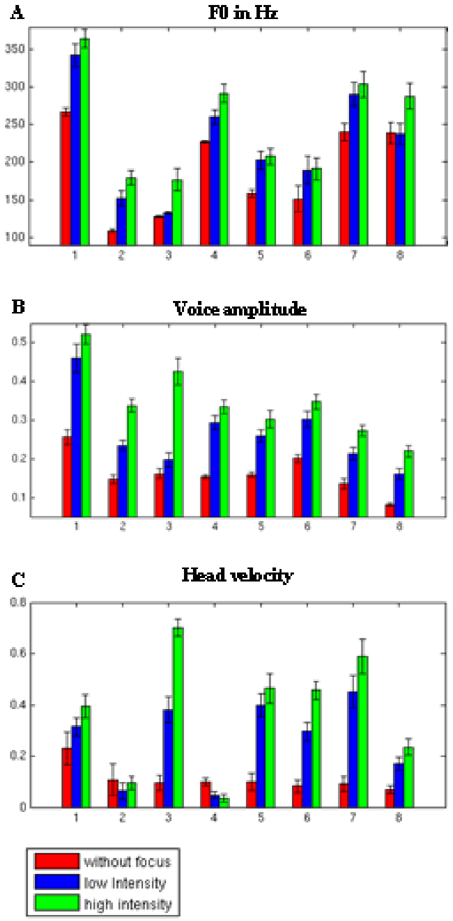

Figure 1: Peak values of the physically analyzed parameters for each speaker across sentences. The error bar depicts the standard error of the mean.

Overall, speakers were able to produce different intensity levels. Interestingly, the specific use of the semiotic channels differed across individuals. Figure 1 shows the mean of the peak values for each of the eight speakers.

3.1. The fundamental frequency f0

The f0 pitch contour was analyzed using the program Praat[1]. The peak frequency of the f0 at the location of the prominence is depicted in Figure 1A. To determine the statistical significance of the variations in the f0, an analysis of variance (ANOVA) was performed with the within-participant factors Sentence, Syntactic Category, and Intensity. Each of the three main factors was found to be significant(all F’s(2,14)>5.5; all p’s<0.02), as were all interactions (all F’s(4,28)>5.1; all p’s<0.01). The Sentence effect means some sentences were emphasized more than others. The Syntactic effect will be addressed more deeply in Section 3.4. The Intensity effect clearly shows that the f0 differs as a function of the intended emphasis intensity. Closer examination of the specific pattern of f0 shows that, for everyone except Speaker 8, the f0 pitch was significantly lower for sentences

[1] Praat (2007), $http://www.fon.hum.uva.nl/praat/$

with No emphasis than those with Low emphasis ($t(1,9)>5.0$; $p<0.05$). Furthermore, The f0 pitch was significantly lower in the Low condition than the High condition for Speakers 1, 2, 3, 4, and 8 ($t(1,9)>9.0$; $p<0.01$). In sum, the f0 does seem to be able to define differences in emphasis intensity, at least for large intensity differences.

3.2. Voice amplitude value

The sound amplitude values in Figure 1B show a straight forward relationship with the different intensity levels. All speakers produced a clear difference in amplitude between each of the three intensity levels (No<Low and Low<High, all t's$(1,9)>5.0$; all p's<0.05).

3.3. Visual information

For the visual analyses, we focused on rigid head motion. By tracking the green markers on the speaker's hat and calculated the square root of the position difference between frame t_n and t_{n+1}, we were able to determine the speaker's head velocity at each frame. For the Low and High conditions, the maximum velocity at the location of the focus was determined. For the No condition, the mean head velocity for the whole sentence was calculated, as this represents the typical head motion behavior of the speaker during normal speaking. The means of the maximum velocity values for each speaker are shown in Figure 1C.

The ANOVA found a significant main effect for Intensity ($F(2,14)=26.7$; $p<0.001$) and Syntactic Category ($F(2.14)=3.9$; $p<0.05$), but not for Sentences ($F(9,63)=1.4$; $p>0.23$). Only the Syntactic Category by Intensity interaction was significant ($F(4,28)=2.8$; $p<0.05$). In other words, head motion reflected differences in emphasis as a function of intensity and syntactic category just as the two acoustic semiotic channels. This will be discussed in more detail in Section 3.4. Unlike the acoustic channels, head motion was relatively consistent across sentences.

As can be seen in Figure 1C, there is considerable variation in the pattern of head motion across the different speakers. In general, head motion was used by 6 of the 8 speakers to show at least some difference between the intensity levels. Speakers 2 and 4 did not show any head motion. Speakers 1 and 5 showed some difference, but only between two intensity levels. Speakers 3, 6, 7, and 8 showed clear distinctions between all the three intensity levels.

3.4. Syntactic categories

As mentioned, the location and syntactic category (adjective, noun, or verb) of the emphasized word was systematically varied. All three semiotic channels showed a significant variation as a function of syntactic category. Here, we examine these differences in more detail. The results are shown in Table 2.

A quick glance at Table 2 shows two important things. First, while the differences as a function of syntactic category can be found in each channel, most of the significant variation lies with changes in voice amplitude. Second, there are strong individual differences. For example, Speaker 5 show no category effect, while Speaker 6 distinguished the three categories in all semiotic channels at all intensity levels. Interestingly, in all cases, the verb was emphasized the strongest. This effect might be due to the syntactic role of the verb or to the fact that, of the three syntactic categories used here, the verb is always the earliest in a sentence.

	F0		Voice Amp		Head	
Spr	low	high	low	high	low	high
1	*	n.s.	n.s.	n.s.	n.s.	*
2	n.s.	n.s.	**	*	n.s.	n.s.
3	***	n.s.	*	**	*	n.s.
4	*	**	*	*	n.s.	n.s.
5	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
6	**	**	**	*	***	***
7	n.s.	n.s.	**	n.s.	n.s.	n.s.
8	n.s.	n.s.	*	***	n.s.	n.s.

Table 2: Statistical analysis of the three syntactic categories (adjective, noun, verb) for each speaker separated into the three semiotic channels and the higher two intensities. The asterisk represents statistically significant differences in the *production* of the three words ("n.s." stands for not significant).

3.5. Discussion

These eight speakers were clearly able to produced different levels of accentuation. Additionally, each of the three semiotic channels (f0 pitch, voice amplitude, and head motion) were used by at least one speaker. Interestingly, there were strong individual differences in the specific distribution of information across the three channels. Finally, verbs seemed to be stressed more than other syntactical categories.

4. Experiment 1

It is now clear that these speakers were able to *produce* different levels of emphasis intensity. In the first experiment, we examined whether or not observers can *perceive* these differences. To do this, we showed participants the sentences and asked them to rate the emphasis intensity using a 7 point Likert scale [14].

4.1. Method

4.1.1. Apparatus

The video sequences were reduced to 256x192 pixels and were presented on a 21" monitor. The participants sat in a dimly lit room in front of the computer screen at a distance of approximately 0.5 meters. The sound was presented to the participants with noise canceling headphones. Responses were entered via a computer keyboard.

4.1.2. Design

For each trial, participants were presented with a black screen and had to press the space bar to start the video sequence. After the sequence ended, the participants were asked to rate the intensity of the emphasis in the sentence using a 7 point scale with a rating of 1 representing a weak intensity and 7 a strong intensity. Note that since each recording only had one emphasis location, this is equivalent to asking the participants to rate the intensity of the emphasized word. Once the response was entered, the request to press the space bar to start the next trial was displayed. The 560 sequences (8 speakers, 10 sentences, 3 intensity levels, and 3 emphasis locations) were shown to each participant in a random order.

4.1.3. Participants

Ten undergraduate students participated in the experiment, all of whom were naive to the purposes of the experiment. They were paid for their participation.

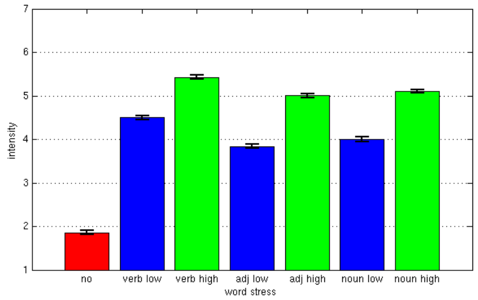

Figure 2: Ratings of the participants for the different intensities and the syntactic categories. Error bar depicts the standard error of the mean.

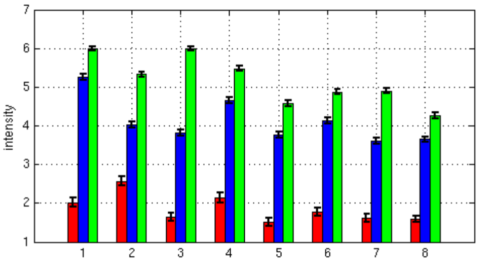

Figure 3: Ratings of the participants shown for each speaker. Error bar depicts the standard error of the mean.

4.2. Results and discussion

The three-way ANOVA found significant main effects for Speaker (F(7,63)=34.5; p<0.001), Syntactic Category (F(9,81)=9.3; p<0.001), and Intensity (F(6,54)=101.3; p<0.001). This shows that some speakers, categories, and intensity levels were rated as being more intense than others. The Speaker by Category and Speaker by Intensity interactions were significant (F(63,567)=3.8; p<0.001 and F(42,378)=13.3; p<0.001, respectively), indicating that the ability to do produce different intensities and the preferential emphasis of certain syntactic categories was different for different speaker. There was also a significant interaction between Intensity and Category (F(54,486)=3.2; p<0.001). The three-way interaction was also significant (F(378,3402)=2.9, p<0.001).

4.2.1. Intensity ratings

Overall, the ratings depicted in Figure 2 clearly varied as a function of intensity, i.e. the High intensity sentences were ranked higher (5.2, on average) than the Low intensity sentences (4.1, on average), which were in turn higher than the No intensity sentences (1.9). As can be seen in Figure 3 the pattern of ratings is remarkably similar across Speakers, indicating that regardless of how the speakers varied emphasis intensity, the different intensity levels were easily and consistently perceived.

Some of the ratings patterns show an obvious similarity to patterns found in the physical analyses. For example, the Low intensity sentences was generally closer to the High intensity than the No intensity sentences in both the ratings and the physical variations. For several speakers (2, 3, and 8), the ratings of the three intensity levels were evenly spaced. The physical correlate of this rating pattern is different for the three speakers: Speaker 2 shows the pattern in both acoustic channels, Speaker 3 shows the pattern only in the head motion profile, and Speaker 8 shows the pattern in the head motion and the voice amplitude.

To further examine the relationship between semiotic channel and intensity ratings we performed a multiple hierarchical regression analysis. The results show that the voice amplitude correlates best with the ratings (R^2 between 0.1 and 0.35) for all speakers except for Speaker 4. Speakers 1 and 2 also showed a correlation with the f0 pitch (R^2 between 0.2 and 0.34). For Speakers 3, 6, and 8, the ratings also correlate with the head motion (R^2 between 0.2 and 0.47). For Speaker 7 the ratings correlates only with the voice amplitude (R^2= 0.21). Although the voice amplitude correlated with the ratings for nearly all speakers, if any other channel also showed a correlation, the R^2 of the voice amplitude was always lower.

In sum, the speakers were able to produce different levels of emphasis intensity. The specific usage of the three semiotic channels differed fundamentally across speakers. Some speakers used a single semiotic channel, others used several channels simultaneously, and some used different channels at different intensity levels. Regardless of how the information was conveyed, however, the effect was perceived to be very similar. That is, fundamentally different distributions of emphasis intensity information led to very similar perceived intensity.

4.2.2. Syntactic category effects

For each speaker, syntactic category, and intensity level, we statistically analyzed the intensity ratings and placed the results in Table 3. Consistent with the physical analyses, most of the speakers were perceived as placing different emphasis intensity on the different syntactic categories, with verb being rated the most intense. In comparing the ratings with the physical analyses, we see some similarities. Trivially, Speaker 5's physical analyses did show a category effect and this is reflected in the intensity ratings. Speaker 6 showed a category effect in every semiotic channel, and there is a clear category effect in the ratings. Speakers 2, 3, and 8 showed a category effect in voice amplitude and in the ratings. In contrast, Speaker 4 clearly showed a category effect in both acoustics channels, but no category effect was found in the ratings.

4.3. Discussion

Several naive observers were asked to rate the perceived intensity of the different sentences. The results show that the observers were able to clearly and easily distinguish between the intensity levels. Moreover, similar rating patterns were obtained regardless of how the individual speakers conveyed emphasis intensity. Both the physical analyses and the perceptual ratings show a very consistent effect of syntactic category. The different speakers place more emphasis on verbs than on the other categories (although this may simply be due to the fact that the verbs always came earlier in the sentence than the other categories), and observers can detect this. It seems that despite constant exposure to the differential intensity ratings of verbs, observers do not automatically compensate, which would have produced equal ratings across all categories. Finally, it is important to mention that the pattern of intensity ratings was the same for each of the ten sentences, suggesting that category effect was not due to the semantics of the individual words.

5. Experiment 2

Experiment 1 showed that fundamentally different patterns of semiotic channel usage lead to remarkably similar patterns of perceived intensity. This naturally raises the question of how

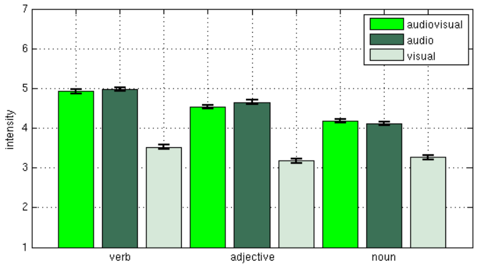

Figure 4: Ratings of the participants for the different presentations split by the syntactic category. Error bar depicts the standard error of the mean.

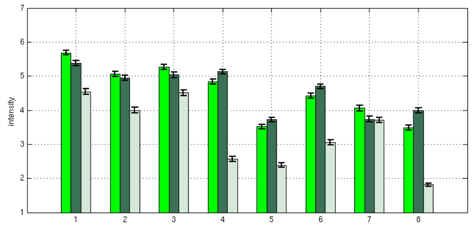

Figure 5: Ratings of the different presentation styles for each speaker. Error bar depicts the standard error of the mean.

the information in the different semiotic channels is perceived and integrated. To address this issue, we used the high intensity sentences of the previous experiment using one of three different presentation styles: Visual, Audio, or Audiovisual.

5.1. Method

Ten new, naive undergraduate students took part in this experiment for financial compensation at standard rates. The same equipment and design as in Experiment 1 were used. The stimulus set consisted of all high intensity sentences. Crossing eight Speakers, ten sentences, three Syntactic Categories (verb, adjective, and noun) with three different presentation styles yielded a total of 720 trials.

5.2. Results and discussion

The results were subjected to a four-way ANOVA with Speaker, Sentence, Syntactic Category, and Presentation Style as within-participant factors. All main effects and all interactions were significant (all F's>10.0, p's<0.001).

5.2.1. Intensity ratings

Overall, as can be seen in Figure 4, ratings in the Audiovisual and Audio conditions do not differ significantly from each other (for all syntactic categories: all t's(1,9)<0.6; n.s.). Ratings in the Visual condition were significantly lower than the two other conditions (all t's(1,9)>14; all p's<0.001). Thus, while there is information in the visual modality (the ratings are well above 1), this information is substantially weaker or less salient than the auditory information, at least when presented in isolation.

Figure 5 shows the ratings for each Speaker and each Presentation Style. In contrast to the overall values, here it can be seen that the Audiovisual and Audio conditions were rated as being different for every Speaker (all t's(1,9)>11.5; all p's<0.001) except for Speaker 2 (t(1,9)=2.6; n.s.). This apparent contradiction can be explained by the fact that some speakers show a lower rating for the audiovisual condition than for the audio only condition (Speakers 4, 5, 6, and 8), while the rest showed the opposite effect. Thus, when the results are averaged across speakers, it looks like there is no difference between the two conditions.

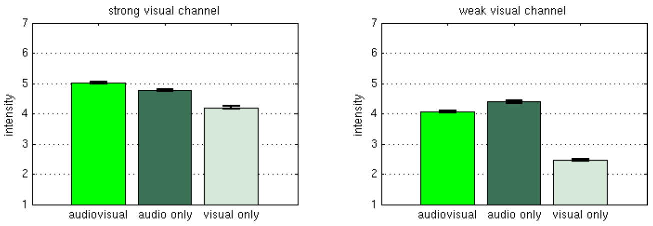

Figure 6: Ratings for the different presentation styles collected for speakers with a strong and a weak visual channel. Error bar depicts the standard error of the mean.

A closer glance at Figure 5 shows that the speakers that were rated lower in Audiovisual than in Audio also seem to have a reasonably low rating on the Visual condition. To further explore this, the data from these speakers four speakers were collected into one group (which will be referred to as the ”weak visual” group). The data from the remaining Speakers form a ”strong visual” group. The averages of these two groups can be found in Figure 6. For the strong visual group, the ratings in the Audiovisual condition were significantly *higher* than for the Audio condition ratings (t(1,9)=87.8; p<0.001). For the weak visual group, the Audiovisual ratings were significantly *lower* than Audio ratings (t(1,9)=53.1; p<0.001). This suggests that weak visual information for emphasis can decrease the overall perceived intensity when combined with audio information. Note that the results cannot be explained by a straightforward averaging of the intensities[2].

Interestingly, only some of the ratings in the Visual condition would have been predicted by the physical analysis. For example, Speakers 5 and 6 showed considerable usage of head motion to specify emphasis, yet received low ratings in the Visual condition. In contrast, Speaker 2 did not seem to use head velocity to specify emphasis, yet received rather high ratings in the Visual condition. The rating of the visual channel, then, does not seems to be not solely related to the head motion. Most likely the visual information is, at least in part, somewhere else such as eyebrow motion.

5.2.2. Syntactic category effects

Overall, the specific pattern of category effects is consistent with the results of the previous experiment. As can be seen in Table 3, an effect of syntactic categories can be found consistently in the Audio and Audiovisual conditions, but rarely in the Visual condition. This suggests that the differential emphasis of verbs is largely an acoustic phenomenon.

5.3. Discussion

Experiment 2 examined the uni- (audio only, visual only) and multimodal (audiovisual) perception of emphasis. While visual information was consistently seen as being the weakest way of conveying emphasis, it is clearly possible to specify emphasis using only visual information. The combination of strong vi-

[2] If, however, the contribution from the visual channel were first subtracted from some expected mean value, which lies between the “weak” and “strong” groups, then a weighted average could explain the results.

	Experiment 1		Experiment 2		
	low	high	Audiovisual	Audio only	Visual only
1	***	n.s.	n.s.	**	n.s.
2	***	***	***	***	***
3	***	**	***	**	***
4	n.s.	n.s.	**	***	n.s.
5	n.s.	n.s.	n.s.	n.s.	n.s.
6	***	*	**	**	n.s.
7	***	**	**	**	n.s.
8	**	***	***	*	n.s.

Table 3: Statistical analysis of the three syntactic categories (adjective, noun, verb) for each speaker. For Experiment 1, this is separated into the two intensities Low and High. For Experiment 2, this is separated into the three presentation styles. The asterisk represents statistically relevant differences in the *perception* of the three words.

sual information with acoustic information lead to an stronger emphasis than was found in either modality alone. In contrast, when weak (but still noticeable) visual information is added to acoustic information, the multimodal emphasis was seen as being weaker than the auditory information alone.

6. General discussion

In this study, we investigated the production and perception of emphasis intensity. We recorded German sentences in which the prosodic prominence varied between No, Low, and High intensity. The location and syntactic category of the emphasis was also systematically varied. Physical analyses of the recordings showed that speakers are able to produce different intensity levels. Considerable individual variance was found in the exact manner of expressing emphasis intensity. Some speakers used primarily voice amplitude, others relied on f0 pitch, and yet others on head velocity. Finally, some relied on various combinations of these information channels. Experiment 1 looked at whether the physically present differences in intensity could be detected by normal observers. The results show that not only could the observers easily and readily detect the differences in the intensity levels, but that the different methods of producing intensity were perceived as being similar.

Experiment 2 examined how visual and auditory information are individually perceived, and how they are integrated. Overall, intensity ratings in the audio and audiovisual conditions were generally similar. While the visual information was considerably weaker than the auditory information, it was nonetheless sufficient to perceived some emphasis. The pattern of results also showed that the information from the various semiotic channels is always integrated in the audiovisual presentations. This has strong implications for the production of computer animated figures, which tend to focus strongly on the use of acoustic information and almost completely ignore visual information. Such a strategy requires, at the very least, an exaggerated acoustic signal to compensate for the lack of visual information in order to produce the desired audiovisual effect.

7. Acknowledgments

This work was supported by the European Network of Excellence ENACTIVE[3] (FP6-IST 002114).

8. References

[1] M. Heldner and E. Strangert, "To what extent is perceived focus determined by f0-cues?" *EUROSPEECH-1997*, pp. 875 – 878, 1997.

[2] C. Gussenhoven, B. Repp, A. Rietveld, W. Rump, and J. Terken, "The perceptual prominence of fundamental frequency peaks," *Journal of the Acoustic Society of America*, vol. 102, pp. 3009 – 3022, 1997.

[3] M. Renwick, S. Shattuck-Hufnagel, and Y. Yasinnik, "The timing of speech-accompanying gestures with respect to prosody," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, p. 2397, 2004.

[4] C. Cave, I. Guaitella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, "About the relationship between eyebrow movements and f0 variations," *Proceedings of the ICSLP*, pp. 2175 – 2179, 1996.

[5] P. Keating, M. Baroni, S. Mattys, R. Scarborough, A. Alwan, E. Auer, and L. Bernstein, "Optical phonetics and visual perception of lexical and phrasal stress in english," *Proceedings of the 15th International Congress of Phonetic Sciences*, pp. 2071 – 2074, 2003.

[6] M. Swerts and E. Krahmer, "Congruent and incongruent audiovisual cues to prominence," *Proceedings of Speech Prosody, Nara (Japan)*, pp. 69 – 72, 2004.

[7] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics*, vol. 30, pp. 555 – 568, 2002.

[8] K. Munhall, J. Jones, D. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility," *Psychological Science*, vol. 15, no. 2, pp. 133 – 137, 2004.

[9] C. Gussenhoven, "Phonology of intonation," *GLOT International*, vol. 6, pp. 271 – 284, 2002.

[10] M. Swerts, "The effect of visual beat gestures on prosodic prominence," *Workshop on Visual Prosody in Language Communication, Nijmegen*, 2007.

[11] M. Liberman and J. Pierrehumbert, "Intonational invariance under changes in pitch ranges and length," in *Language Sound Structure*, M. Arono and R. Oehrle, Eds., 1984, pp. 157 – 233.

[12] M. Kleiner, C. Wallraven, M. Breidt, D. Cunningham, and H. Blthoff, "Multi-viewpoint video capture for facial perception research," *Captech 2004 - Workshop on modelling and motion capture techniques for virtual environments*, pp. 55 – 60, 2004.

[13] J. P. de Ruiter, S. Rossignol, L. Vuurpijl, D. Cunningham, and W. J. M. Levelt, "Slot: A research platform for investigating multimodal communication," *Behavior Research Methods, Instruments, & Computers*, vol. 35, no. 3, pp. 408–419, 2003.

[14] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 140, no. 55, 1932.

[3] http://www.enactivenetwork.org