Differences in the audio-visual detection of word prominence from Japanese and English speakers

Martin Heckmann[1]*, Keisuke Nakamura*[2]*, Kazuhiro Nakadai*[2]

[1]Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany
[2]Honda Research Institute Japan Co. Ltd., Wako-shi, Saitama 351-0188, Japan
martin.heckmann@honda-ri.de, keisuke@jp.honda-ri.com, nakadai@jp.honda-ri.com

Abstract

We have previously shown that for English speakers information on the mouth shape of a speaker is a powerful feature for the machine based discrimination of prominent from non-prominent words. In this paper we extend our analysis to data from Japanese speakers. We compare the discrimination performance of the different acoustic and visual features we extract for the two languages. This comparison shows a much wider variability in discrimination scores for the different speakers and the different features in the English dataset than in the Japanese dataset. Despite previous hints that visual speech and word prominence perception by Japanese listeners can yield inferior performance compared to English listeners we see that our discrimination scores are high and very similar for the English and Japanese speakers which indicates that at least the speakers signal prominence with a similar level of consistency in both languages.

Index Terms: prosody, prominence, visual, audio-visual, Japanese

1. Introduction

Speech is not only what we say but also how we say it. Already quite some time ago Thompson realized that this "How", the prosody of speech, can also be seen from the movements of the speaker [1]. Since then quite a few researchers showed the importance of the different facial regions in the perception of prosodic variations [2, 3, 4]. Not only the mouth [5] but also the eye brows [6, 7] and the rigid head movements are important information sources [8, 9]. These perceptual experiments were also supported by results showing mainly stronger and more rapid facial movements for stressed words [10, 11, 12].

Yet it is not only the prosody of the speech which can be perceived from a speaker's face but also the lexical content [13]. The most revealing sign of this is the McGurk effect. When different speech sounds are presented visually and acoustically subjects perceive an intermediate sound which is neither present in the acoustic nor the visual channel [14]. Based on this insight a large variety of systems for the audio-visual recognition of speech were developed which in particular show that the inclusion of the visual channel can improve the recognition scores in difficult situations [15, 16, 17, 18]. The machine detection of prosodic cues is still a challenging issue [19, 20, 21]. Despite the fact that audio-visual emotion recognition is already a well established field of research [22] the system we presented in [23] was to our knowledge the first system to show that the integration of visual information yields to more robust and better results for prosodic analysis, in particular for the determination of word prominence.

It is assumed that the audio-visual integration in speech perception depends on culture and language. In [24, 25, 26] and [27] Japanese, respectively Chinese, listeners showed a weaker McGurk effect than English listeners, at least for uncorrupted acoustic speech. One hypothesis for this weaker effect is that in Asian culture it is much less common to look into a speaker's face than in Western culture [27]. Consequently, also the perception of visual prosodic cues should be much weaker for Asian people than for Westerners. In [28] the audio-visual perception of contrastive focus in French and Japanese listeners was investigated. The results showed that the Japanese listeners were much less able to use the visual information to determine the word in focus. On the other hand these and previous results leave it open if the visual production of speech and prosodic cues might also be weaker in Japanese speakers.

In this paper we use the same system we developed in [29] to discriminate audio-visually prominent from non-prominent words and transfer it to Japanese speakers. Based on the previous results for English speakers and the new results for Japanese speakers we discuss the differences between the two languages in particular in respect to the role the different acoustic and visual cues play. In the next section an overview on the recording of the data will be given. After that Section 3 describes the different features extracted from the acoustic and visual channel. Following this Section 4 will present the results of the classification experiments. Then we will discuss the results in Section 5 and give a conclusion in Section 6.

2. Dataset

To elicit prosodic variations we designed a small interactive game which participants were playing with a computer. The target of the game was to assemble a cartoon out of tiles. Thereby participants controlled the computer via speech and recordings of the interaction were made with a microphone and a camera. Subjects were sitting in front of the computer and were not particularly restricted in their movements. The microphone was mounted above the computer screen and the camera behind the screen (compare Fig. 1). No visual markers were attached to the subjects. We implemented this game in English and Japanese. An exemplary utterance during the game would be "Place green in A two" in English and "Midori wo A no ni ban ni oite kudasai" in Japanese. The interaction with the computer was embedded in a Wizard of Oz experiment where a Wizard controlled the reactions of the system. Occasionally, a misunderstanding of one word of the sequence was triggered and the corresponding word highlighted, verbally and visually. Verbal feedback was based on the FESTIVAL speech synthesis system [30]. The subjects were told to repeat in these cases the phrase as they

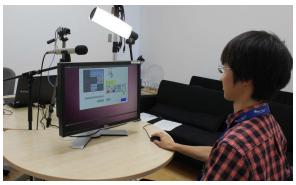

Figure 1: The setup of the recording of the Japanese dataset

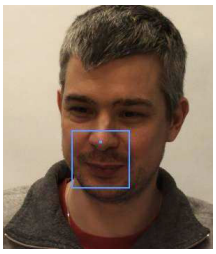

Figure 2: Recorded image after cropping to face region, nose detection, downsampling and highlighting of the mouth region.

would do with a human, i. e. emphasizing the previously misunderstood word. However, they were not allowed to deviate from the sentence grammar by e. g. beginning with 'No'. This was expected to create a narrow focus condition (in contrast to the broad focus condition of the original utterance) and thereby making the corrected word highly prominent.

We recorded 16 subjects for the English dataset. Eight females and eight males, eight speaking British English as their sole native language, three being bilingual British English/German, four speaking American English as their sole native language and one being bilingual American English/German were recorded. Bilingual subjects were raised in Germany but had one native English parent. For the Japanese dataset we recorded 15 subjects all having Japanese as their sole native language.

The audio signal was originally sampled at 48 kHz and later downsampled to 16 kHz. For the video images a CCD camera with a resolution of 1280×1024 pixel and a frame rate of 25 Hz was used. The identical microphone and the same camera model were used for recording both datasets. We used a forced alignment to determine the boundaries of the different words. For the English dataset we used HTK [31] for the alignment and for the Japanese data Julius [32].

For further processing in both databases those turns where the original utterance and a correction were available were selected. Overall we have 2683 turn pairs (original utterance + correction), i. e. on average ≈ 160 turn pairs per speaker for the English dataset and 1931, i. e. on average ≈ 130 turn pairs per speaker, for the Japanese dataset. From these the word which was emphasized in the correction was determined. Then it was extracted as well in the original utterance as in the correction. This yields a dataset with each individual word taken from a broad and a narrow focus condition.

3. Features

To extract word prominence from the audio-visual signals we used the features described in [23, 29]. They are briefly explained in Table 1. The fundamental frequency is extracted according to [33]. For the extraction of visual features we used the openCV library to detect the face and then the nose in the face region [34]. We use the nose position to extract the rigid head movements. Furthermore, as the nose position in the face is only slightly affected by the articulation process we also use it to determine the position of the mouth region. Starting from the nose position we use a fixed and for all speakers identical offset to determine the mouth region. After downsampling by a factor of 2 this yields an image of 100×100 pixels of the mouth region (compare Figure 2). On these images a two-dimensional Discrete Cosine Transform (DCT) was calculated. Out of the 10000 coefficients per image the 50 with the highest energy were selected. This was done by calculating for each speaker separately the mean energy of all 10000 coefficients on a randomly selected subset of 10% of the data. Consequently we obtain 50 coefficients per frame to capture the mouth shape. From these features (where appropriate) the mean value for each word was calculated and used in the subsequent analysis. The beginning and end of the word was taken from the forced alignment. All visual features, i. e. for the nose and the mouth shape, were smoothed along the time axis with a 5-th order FIR lowpass filter with a cut-off frequency of 5 Hz. Furthermore, first and second derivatives (Δ and $\Delta\Delta$) were calculated.

e	energy of the word relative to the mean of the utterance
D_W	duration of the word
D_G	sum of the duration of the gap before and after the word
D	combination of D_W and D_G
f_0^M	maximal value of the parabolic approximation of the fundamental frequency of the word, i. e. value at vertex
f_0^c	maximal curvature of the parabolic approximation of the fundamental frequency of the word, i. e. curvature at vertex
$\bar{f}_0$	mean fundamental frequency of the word relative to the mean of the utterance
f_0	combination of f_0^M, f_0^C and $\bar{f}_0$
DCT	50 DCT coefficients extracted from the mouth region with the highest energy
y	nose y position relative to the mean of the utterance
$y\Delta$, $y\Delta\Delta$	nose y velocity and acceleration

Table 1: Features used to extract word prominence. For details see [23, 29].

4. Results

In the same way as in [29] we trained a Support Vector Machine (SVM) with a Radial Basis Function Kernel to discriminate prominent from non-prominent words using the LibSVM library [35]. While doing so we performed for each feature combination a grid search for C, the penalty parameter of the error term, and γ, the variance scaling factor of the basis function, using the whole dataset. Prior to the grid search the data was normalized to the range $[-1 \dots 1]$. Using the found optimal parameters the SVM was trained on 75% of the data and tested on the remaining 25%. Hereby a 30 fold cross validation in which the data set was always split such that an identical number of elements is taken from both classes was run. To establish the 30 sets a sampling with replacement strategy was applied. This process was performed individually for each speaker in each dataset.

4.1. Audio results

In Table 2 the results of the individual features are given. The results are averaged over all speakers of the respective dataset. We saw already previously that for the English dataset the combination of the duration of the word and the gap before and after

	e	D_W	D_G	D	SE	$\bar{f}_0$	f_0^M	f_0^c	f_0	DCT	y	$y\Delta$	$y\Delta\Delta$	DCT+ y	DCT+ $y\Delta$	DCT+ $y\Delta\Delta$	DCT+ $y{+}y\Delta{+}y\Delta\Delta$
E	58.6	62.8	59.0	64.7	58.8	62.6	65.2	64.8	68.2	66.9	54.9	57.5	57.1	67.0	67.7	67.4	67.8
	(3.6)	(6.9)	(6.2)	(9.4)	(8.3)	(9.0)	(8.9)	(5.7)	(8.9)	(8.9)	(6.6)	(8.7)	(8.9)	(9.0)	(9.1)	(9.1)	(9.3)
JP	61.2	62.9	56.4	66.5	59.9	63.7	62.0	55.1	66.5	65.8	54.6	57.6	55.0	65.5	66.8	66.0	66.6
	(4.4)	(6.0)	(6.5)	(7.1)	(5.6)	(6.2)	(4.5)	(4.1)	(6.3)	(5.5)	(5.4)	(7.2)	(6.4)	(5.5)	(6.5)	(6.0)	(5.8)

	e+D	e+D+DCT	e+D+SE	e+D+SE+DCT	e+D+SE+f_0	e+D+SE+f_0+DCT	e+D+SE+f_0+DCT+$y\Delta$
E	69.5	72.4	71.6	73.7	74.9	74.9	75.0
	(8.5)	(8.5)	(9.4)	(9.1)	(11.0)	(11.1)	(11.1)
JP	68.4	70.9	70.3	72.5	72.2	72.9	73.4
	(7.8)	(7.3)	(7.5)	(7.7)	(8.1)	(7.7)	(7.8)

Table 2: Classification scores in % averaged over all speakers in the respective dataset. See Table 1 for the abbreviations used.

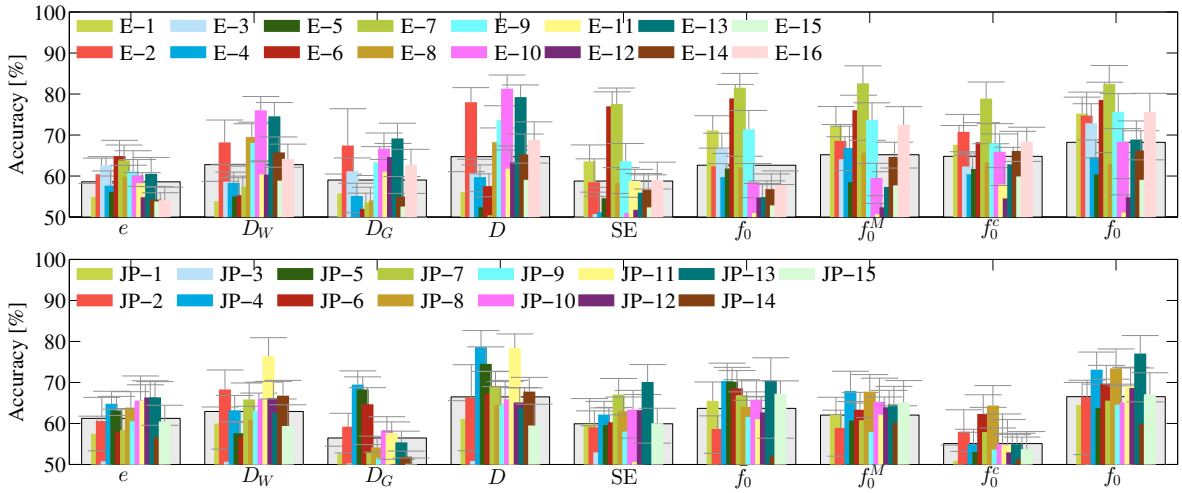

Figure 3: Discrimination accuracies for different acoustic features and feature combinations for each speaker (English in top and Japanese in bottom figure). The grey bars in the background visualizes the average over all speakers for a given feature or feature combination. The short horizontal lines indicate the standard deviation of the 30 fold cross validation. See Table 1 for an explanation of the abbreviations.

the word yields better results than either feature alone [29]. Yet for the Japanese speakers the improvement from this combination is notably stronger than for the English speakers. Also for the Japanese speakers the combination feature f_0 consisting of the fundamental frequency value at the vertex of the parabola approximation f_0^{M} with the curvature at the vertex f_0^{Curv} and the mean of the fundamental frequency in the word relative to the utterance $\bar{f}_0$ also leads to a significant improvement. However, the results for the feature capturing the curvature of the fundamental frequency f_0^{Curv} are much lower for the Japanese speakers than for the English speakers. On the other hand the relative energy of the word is a stronger feature for the Japanese speakers than for the English speakers.

In Fig. 3 the results for each individual speaker are depicted. The ordering of the English speakers is E-1 ... 6 are British English male, E-7 ... 11 British English female, E-12 and 13 US English male and E-14 ... 16 US English female. As can be seen the variation of the results from speaker to speaker is much larger for the English dataset. To further quantify this we also added the standard deviation of the results for each speaker for a given feature in Table 2. The standard deviation is particularly high for the English speakers for the mean of the fundamental frequency in the word relative to the utterance $\bar{f}_0$, the fundamental frequency value at the vertex of the parabola approximation f_0^{M} and the duration of the word D_W.

4.2. Visual results

When looking at the results for the visual features in Table 2 we see that the DCT feature yields very similar results for both datasets. In particular these results are better than any individual acoustic feature and only slightly inferior to those of the combined fundamental frequency feature f_0 and in the Japanese case also the combined duration feature D. We also calculated FFT and PCA features in a similar way but they always led to inferior performance as the DCT. Consequently, we did not include them in the following analysis. When looking at the features derived from the nose dynamics we see that they are all close to chance level for both datasets. However, a look on the detailed results for each speaker depicted in Fig. 4 reveals that there is a large inter speaker variation in both datasets. In particular the nose speed and acceleration are informative for some speakers. In the English dataset the performance for the DCT (from $56-86\%$ correct) and the nose dynamics (e. g. nose velocity from $52-78\%$ correct) varies very strongly. In the Japanese dataset the performance of the DCT feature is much more homogeneous (from $55-75\%$ correct). This can also be

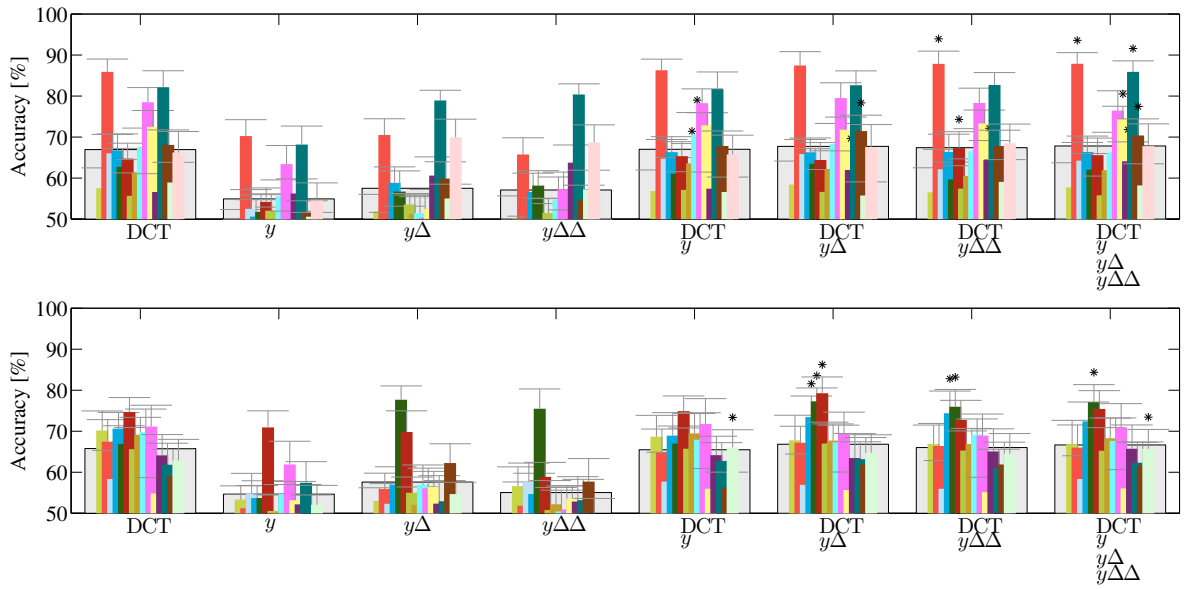

Figure 4: Discrimination accuracies for different visual features and feature combinations for each speaker (English in top and Japanese in bottom figure). The grey bars in the background visualizes the average over all speakers for a given feature or feature combination. The short horizontal lines indicate the standard deviation of the 30 fold cross validation. The asterisk marks results where the combination of visual features are statistically significantly better than the DCT results.See Table 1 for an explanation of the abbreviations.

verified by looking at the standard deviation values in Table 2. Yet for the nose dynamics the variation is similar to that found in the English dataset (e. g. nose velocity from $51 - 79\%$ correct).

4.3. Audio-visual results

Finally, we want to compare the classification results we obtain when we combine different features, i. e. different visual features, different acoustic features and acoustic and visual features. From Fig. 4 we can see that for both datasets combining the DCT feature with the nose dynamics only marginally improves the performance averaged over all speakers. Yet it yields to significant improvements for some speakers (at most 5 for the English and 3 for the Japanese dataset). When combining different acoustic features we see a notable improvement of the results in both datasets (compare Table 2 and Fig. 5). Adding the visual DCT feature to these different acoustic feature combinations improves performance as long as the f_0 feature is not used. When combining all acoustic features, including f_0, we obtain an average performance of 75% for the English data and 72% for the Japanese data. Additionally including the DCT feature does not change the performance for the English data but improves it to 73% for the Japanese data. As we would expect from the large variation of the results from speaker to speaker things look a bit different when looking at individual speakers. Here we see that for some speakers there is quite a notable improvement from combining all acoustic and visual features. There are 5 speakers in the English and 6 in the Japanese dataset which show such an improvement. If we now also add the nose velocity as a feature there is in both datasets one speaker for which the performance is further improved.

5. Discussion

When comparing the English and the Japanese dataset we first observe that the results are significantly more homogeneous for the different speakers in the Japanese dataset. One reason for this might be that the group of participants in the Japanese dataset was more homogeneous (mean age 27.9, stdv. 7.8). Eleven of the fifteen participants were 4th year university engineering students. In contrast the English speakers had a wide age range (18-60 years, mean 36.4, stdv. 11.0), different language background, and different professions. After the experiment we also handed out a questionnaire to the subjects. The results confirmed the impression during the experiment that the Japanese subjects in contrast to the English subjects perceived it much less as a playful game and more like a test. On a 5 point scale from -2 to 2 the English and Japanese subjects considered the interaction with the system as 0.27 and -0.08 natural, emphasizing the corrected word as 0.47 and 0.75 natural and the game overall as 0.53 and 0.92 difficult, with 2 extremely easy. From this we conclude they considered the interaction overall as quite natural and not difficult. Regarding stress English subjects reported -1.2 and Japanese subjects 0.43 with -2 ”totally relaxed” and 2 ”afraid not to make it right”. Also asked for how tired they were (-2 ”very exhausted” and 2 ”totally fresh”) English subjects reported 0 and Japanese subjects -0.53. This indicates that the Japanese subjects felt a much higher tension during the game and hence might have not spoken as freely as their English counterparts. In general Japanese are reported to feel a high degree of moral and social obligation [36].

Another reason for this larger homogeneity of the results for the Japanese speakers could also be cultural aspects in terms of language usage and showing emotions. For example it has been observed that Japanese people display negative emotions very differently when they thing they are not observed compared to when they are observed, i. e. they hide them with a smile [37].

When looking on the average discrimination accuracies we see smaller differences between the English and the Japanese datasets. Overall the results were slightly better for the English speakers. We see a clear difference in the way fundamental frequency is used. For the English speakers we see the expected quick raise and quick drop of the fundamental frequency inside the prominent word [38]. This can bee seen from the high accu-

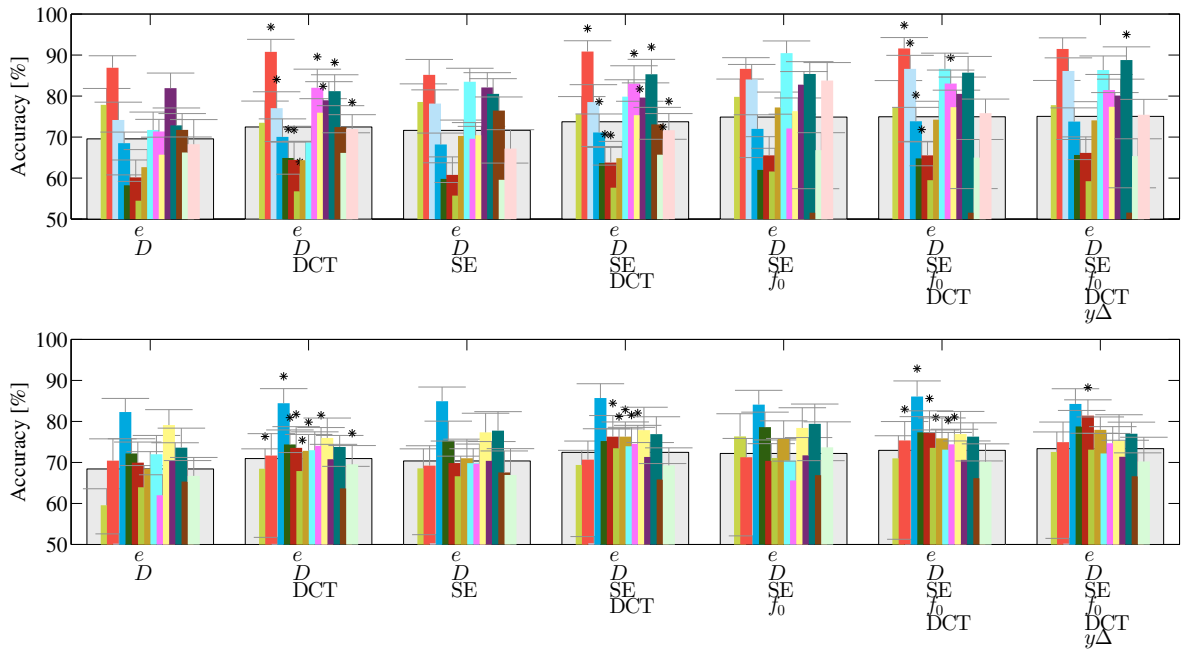

Figure 5: Discrimination accuracies for different acoustic and audio-visual feature combinations for each speaker(English in top and Japanese in bottom figure). The grey bars in the background visualizes the average over all speakers for a given feature or feature combination. The short horizontal lines indicate the standard deviation of the 30 fold cross validation. The asterisk marks results where the combination of acoustic and visual features are statistically significantly better than the corresponding acoustic only results or, as for the last values, as the results without the nose y velocity. See Table 1 for an explanation of the abbreviations.

racies obtained by the fundamental frequency curvature feature f_0^c. Due to the different way prominence is marked in Japanese we do not see this for the Japanese speakers. In Japanese pitch accent has a lexical role. As a consequence their is an interaction between the lexical pitch accent and prominence marking via pitch [39, 40]. Prominence is mainly marked by an increase in pitch range, i. e. the fundamental frequency inside the word is raised and that of the neighboring words reduced. In our data this can be seen from the $\bar{f}_0$ feature capturing the mean fundamental frequency of the word relative to the utterance. $\bar{f}_0$ shows higher discrimination scores for the Japanese speakers than for the English speakers.

In the English dataset there are some speakers which rather use duration (E-2, E-8, E-10, E-13) to signal prominence and others fundamental frequency (E-1, E-3, E-6, E-7). This repartition is independent on language background and gender. In case of the Japanese speakers the scores for fundamental frequency and also to a lesser extent for duration are rather homogeneous over all speakers and one can not easily divide speakers into different sets based on these criteria.

When looking at the results for the visual features averaged over all speakers we see little differences between the English and Japanese dataset. In both cases the DCT feature yields classification scores similar to the individual acoustic features. When looking at the individual speakers the two datasets differ notably. Again, for the Japanese speakers the results are much more homogeneous and do not surpass 75% correct for any speaker. On the other hand for the English dataset we see three speakers with more than 75% correct from the DCT feature, with the best 86% correct. Hence there is a wider spread for the English speakers and there are a few which visually signal prominence very clearly. The nose dynamics show mixed results in both datasets. On average they are not informative but there are in both datasets speakers where the discrimination accuracy reaches almost 80% correct. When comparing our results for the visual discrimination of prominent from non-prominent words to previous findings on the audio-visual speech perception by Japanese listeners and in particular their weaker ability to use the visual information to recognize the word [24, 25, 26] or to identify the prominent word [28] we conclude that this effect can not be linked to a weaker or less consistent production of such cues by Japanese speakers.

Averaged over all speakers the combination of acoustic and visual features does not yield better results for both datasets. Yet for individual speakers such a combination improves the discrimination accuracies quite notably as well for the English as the Japanese dataset. In both datasets there is one speaker where the nose y-velocity caries enough information that it improves the accuracy when it is added to all acoustic features and the DCT feature.

Overall one can say that the visual channel is quite informative to discriminate prominent from non-prominent words as well in English as in Japanese. For the English speakers there seems to be a wider range of different strategies employed to signal prominence and also the consistency and markedness of the signals seems to vary much more than for the Japanese speakers.

6. Conclusion

We previously developed a system to discriminate prominent from non-prominent words based on audio-visual information. In this paper we also included results from Japanese data into the evaluation and compared them with those we obtained from the English data. In this comparison we saw a much higher homogeneity in the results for the Japanese speakers. We attribute this to a more homogeneous set of speakers, a higher tension felt by the Japanese subjects during the experiment and cultural

aspects.

Previous research on audio-visual speech perception and word prominence by Japanese indicated that they use visual information to a lesser extent than English listeners do. Our results show that the discrimination accuracies we obtain for the Japanese speakers are very similar on average to those of the English speakers. From this we conclude that even if the Japanese listeners might not use the visual information as well as English listeners do at least the Japanese speakers produce it at a similar level of consistency as their English counterparts.

7. Acknowledgments

We want to thank Petra Wagner, Britta Wrede and Heiko Wersing for fruitful discussions. Furthermore, we are very grateful to Rujiao Yan and Samuel Kevin Ngouoko for helping in setting up the visual processing and the forced alignment, respectively. Many thanks to Mark Dunn for support with the cameras and the recording system as well to Mathias Franzius for support with tuning the SVMs and Merikan Koyun for help in the data preparation and H. Nakamura for his support in the recording of the Japanese subjects. Special thanks go to our subjects for their patience and effort.

8. References

[1] Dorothy Mossman Thompson, "On the detection of emphasis in spoken sentences by means of visual, tactual, and visual-tactual cues," *The Journal of General Psychology*, vol. 11, no. 1, pp. 160–172, 1934.

[2] Lynne E Bernstein, Silvio P Eberhardt, and Marilyn E Demorest, "Single-channel vibrotactile supplements to visual perception of intonation and stress," *J. Acoust. Soc. Am.*, vol. 85, pp. 397, 1989.

[3] H.P. Graf, E. Cosatto, V. Strom, and F.J. Huang, "Visual prosody: Facial movements accompanying speech," in *Int. Conf. on Automatic Face and Gesture Recognition*. IEEE, 2002, pp. 396–401.

[4] J. Beskow, B. Granström, and D. House, "Visual correlates to prominence in several expressive modes," in *Proc. INTERSPEECH*. ISCA, 2006, pp. 1272–1275.

[5] M. Swerts and E. Krahmer, "Facial expression and prosodic prominence: Effects of modality and facial area," *Journal of Phonetics*, vol. 36, no. 2, pp. 219–238, 2008.

[6] E. Krahmer and M. Swerts, "More about brows," *From brows to trust*, pp. 191–216, 2005.

[7] M.L. Flecha-García, "Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in english," *Speech Communication*, vol. 52, no. 6, pp. 542–554, 2010.

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

[9] S. Al Moubayed and J. Beskow, "Effects of visual prominence cues on speech intelligibility," in *Proc. Int. Conf. Auditory Visual Speech Process. (AVSP)*. ISCA, 2009, vol. 9, p. 16.

[10] Rebecca Scarborough, Patricia Keating, Sven L Mattys, Taehong Cho, and Abeer Alwan, "Optical phonetics and visual perception of lexical and phrasal stress in english," *Language and Speech*, vol. 52, no. 2-3, pp. 135–175, 2009.

[11] M. Dohen, H. Lœvenbruck, H. Harold, et al., "Visual correlates of prosodic contrastive focus in french: Description and inter-speaker variability," in *Speech Prosody*, Dresden, Germany, 2006.

[12] E. Cvejic, J. Kim, C. Davis, and G. Gibert, "Prosody for the eyes: Quantifying visual prosody using guided principal component analysis," in *Proc. INTERSPEECH*. ISCA, 2010.

[13] W.H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 26, pp. 212, 1954.

[14] H McGurk and J MacDonald, "Hearing lips and seeing voices.," *Nature*, vol. 264, no. 5588, pp. 746, 1976.

[15] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[16] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Applied Signal Process.*, vol. 11, pp. 1260–1273, 2002.

[17] D. Kolossa, S. Zeiler, A. Vorwerk, and R. Orglmeister, "Audio-visual speech recognition with missing or unreliable data," in *Proc.Int. Conf. Auditory Visual Speech Process. (AVSP)*, 2009.

[18] T. Yoshida, K. Nakadai, and H.G. Okuno, "Automatic speech recognition improved by two-layered audio-visual integration for robot audition," in *Proc. 9th IEEE-RAS Int. Conf. on Humanoid Robots*. IEEE, 2009, pp. 604–609.

[19] M.Q. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," *Computer Speech & Language*, vol. 6, no. 2, pp. 175–196, 1992.

[20] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[21] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Proc. EUROSPEECH*. ISCA, 2005.

[22] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic, "Avec 2011–the first international audio/visual emotion challenge," pp. 415–424, 2011.

[23] M. Heckmann, "Audio-visual evaluation and detection of word prominence in a human-machine interaction scenario," in *Proc. INTERSPEECH*, Portland, OR, 2012, ISCA.

[24] Kaoru Sekiyama and Yoh'ichi Tohkura, "Mcgurk effect in non-english listeners: Few visual effects for japanese subjects hearing japanese syllables of high auditory intelligibility," *J. Acoust. Soc. Am.*, vol. 90, pp. 1797, 1991.

[25] Kaoru Sekiyama and Yoh'ichi Tohkura, "Inter-language differences in the influence of visual cues in speech perception.," *Journal of Phonetics*, 1993.

[26] Dominic W Massaro, Michael M Cohen, Antoinette Gesi, Roberto Heredia, et al., "Bimodal speech perception: An examination across languages," *Journal of Phonetics*, vol. 21, pp. 445–478, 1993.

[27] Kaoru Sekiyama, "Cultural and linguistic factors in audiovisual speech processing: The mcgurk effect in chinese subjects," *Perception & Psychophysics*, vol. 59, no. 1, pp. 73–80, 1997.

[28] Marion Dohen, Harold Hill, Chun-Huei Wu, et al., "Auditory-visual perception of prosodic information: Inter-linguistic analysis-contrastive focus in french and japanese," in *Proc. Int. Conf. Auditory-Visual Speech Proc. (AVSP)*, 2008, pp. 89–93.

[29] M. Heckmann, "Inter-speaker variability in audio-visual classification of word prominence," in *submitted to INTERSPEECH*, 2013.

[30] A.W. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," Tech. Rep., 1998.

[31] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University, Cambridge, United Kingdom, 1995.

[32] Akinobu Lee and Tatsuya Kawahara, "Recent development of open-source speech recognition engine julius," in *Proc. Asia-Pacific Signal and Information Proces. Assoc. An. Summit and Conf.*, 2009, pp. 131–137.

[33] M. Heckmann, F. Joublin, and C. Goerick, "Combining rate and place information for robust pitch extraction," in *Proc. INTERSPEECH*, Antwerp, 2007, pp. 2765–2768.

[34] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[35] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[36] William Caudill, "The influence of social structure and culture on human behavior in modern japan," *Ethos*, vol. 1, no. 3, pp. 343–382, 1973.

[37] Wallace V Friesen, "Cultural differences in facial expressions in a social situation: An experimental test on the concept of display rules.," Unpublished doctoral dissertation, University of California, San Francisco, 1973.

[38] Yi Xu and Ching X Xu, "Phonetic realization of focus in english declarative intonation," *Journal of Phonetics*, vol. 33, no. 2, pp. 159–197, 2005.

[39] Haruo Kubozono, "Focus and intonation in japanese: Does focus trigger pitch reset," *Interdisciplinary Studies on Information Structure (Working Papers of the SFB 632)*, 2005.

[40] Jennifer Venditti, Kikuo Maekawa, and Mary E Beckman, *Handbook of Japanese linguistics*, chapter Prominence marking in the Japanese intonation system, pp. 456–512, Oxford University Press, 2008.