Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Audio-visual Evaluation and Detection of Word Prominence in a Human-Machine Interaction Scenario

Martin Ernst Heckmann

2012

Preprint:

This is an accepted article published in Proc. INTERSPEECH. The final authenticated version is available online at: https://doi.org/[DOI not available]

Audio-visual Evaluation and Detection of Word Prominence in a Human-Machine Interaction Scenario

Martin Heckmann

Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany

martin.heckmann@honda-ri.de

Abstract

This paper investigates the audio-visual correlates and the detection of word prominence. Subjects were interacting with a computer in a small game which created a broad and a narrow focus condition. Audio-visual recordings with a distant microphone and without visual markers were made. As acoustic features duration, intensity, fundamental frequency and spectral emphasis were calculated. From the visual channel head movements and image transformation based features from the mouth region were extracted. First the results show that the extracted features are significantly different for the two focus conditions (broad and narrow). Based on classification results it is demonstrated that they can be differentiated without knowledge of the word identity with accuracies of approx. 80%. Furthermore, it is shown that the visual channel by itself yields accuracies notably better than chance (approx. 65%) and that a combination of both modalities increases performance to approx. 85%.

Index Terms: prosody, prominence, visual, audio-visual, spectral emphasis, lip movement, head movement

1. Introduction

Despite the fact that the prosody of speech plays a very important role in human-human communication it is rarely used in spoken dialog systems [1]. Several attempts have been made in the past but fell short of making their way into mainstream [2, 3, 4]. One of the reasons might be that prosodic cues are notoriously difficult to extract reliably.

It is well known that visual speech, i. e. the movements of the speakers mouth and face, improves the accuracies of automatic speech recognition, particular in difficult situations [5, 6, 7, 8]. Furthermore, it has been shown, that the visual channel also conveys relevant information for the perception of prosody [9, 10, 11] and in particular for word prominence [12, 13]. Hence it is promising to investigate how this information can be used for the extraction of prosodic features. So far few studies have attempted to quantify and extract prosodic information from the visual channel [14, 15]. In these cases additional markers in the speakers face were used.

In this paper we will investigate which visual prosodic information can be extracted from images of a speakers face interacting with a computer without additional markers positioned in his face. Furthermore, we will see how this information can be combined with information extracted from the acoustic channel to better discriminate highly prominent from less prominent words.

In the next section an overview on the recording of the data will be given. After that Sec. 2 describes the different features extracted from the acoustic and visual channel. Following this Sec. 3 will present the results of the data analysis and the recognition experiments. In the last section we will discuss the results and give an outlook.

2. Audio-visual Database

As prosody is one of the more subtle characteristics of speech the recording of the data was set up to create a situation which allows subjects to behave naturally. They were told that their task was to play a game with a computer where the goal was to assemble a cartoon out of tiles. This yielded utterances of the form "**put green B one**". Colors were red, blue, green and white, letters varied between A...R and numbers from 1...9. In the Wizard of Oz experiment the system would sometimes misunderstand a (single) word of the user command. It would then give visual and verbal (using the FESTIVAL speech synthesis system [16]) feedback and subjects were told to correct the utterance by repeating it "as they would do with a human" but by strictly keeping to the grammar and not introducing any term, e. g. "no". To further clarify it they were told to emphasize the misunderstood word. Via this setting it was expected to elicit first a broad focus condition and then in the correction a narrow focus condition in which the corrected word will be highly prominent [17]. During the experiment the subjects did not have to wear a headset microphone nor were markers positioned on their body.

In total 3 subjects, 1 female and 2 male, either speaking British English as their sole native language or being bilingual British English/German were recorded. 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. Due to the strong resemblance of the recorded speech in grammar and vocabulary to that of the Grid Corpus [18] a speech recognition system trained on that corpus could be used to perform a forced alignment on the acquired data.

For further processing those turns where the original utterance and a correction were available were selected. This yielded 137 turn pairs (original utterance + correction) for subject A, 146 for subject B and 94 for subject C. 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. The words in the narrow focus condition are expected to be notably more prominent than in the broad focus condition. An informal listening test confirmed this expectation.

3. Features

In the following experiments the features described in table 1 which have previously been proposed to correlate with word

Acoustic	
dur	duration of the word
en	energy relative to the mean of the utterance
f0	mean fundamental frequency (extracted according to [19, 20])
SE	spectral emphasis, i. e. the difference between the overall intensity and the intensity in a dynamically low-pass-filtered signal with a cut-off frequency of 1.5 f0 [21]
Visual	
y	nose y position relative to the mean of the utterance
$\Delta, \Delta\Delta$	first and second derivative

Table 1: Description of the different features.

prominence were used. From these features the mean value for each word was calculated and used in the subsequent analysis (except for duration). The beginning and end of the word was taken from the forced alignment.

For the visual modality the openCV [22] was applied to first detect the face in each image frame and then determine the nose position. As the nose moves only slightly relative to the skull during articulation it yields information on the rigid head movements. In addition to the nose position also image transformations calculated on the mouth region were used as feature. For determining the mouth region from the images a fixed and for all speakers identical offset from the nose was used and also the size of the mouth region was kept identical. After downsampling by a factor of 2 this yields an image of 80×80 pixels of the mouth region which is used for further analysis (compare Fig. 1). On these images either a two-dimensional Fast Fourier

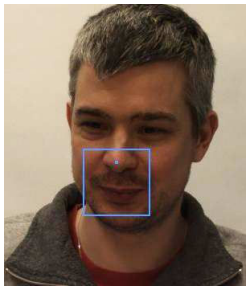

Figure 1: Image from recording after cropping to face region, nose detection, downsampling and highlighting of the mouth region.

Transform (FFT) or Discrete Cosine Transform (DCT) was calculated. Out of the 6400 coefficients per image for FFT or DCT respectively the 50 with the highest energy were selected. This was done by calculating for each speaker separately the mean energy of all 6400 coefficients on a randomly selected subset of 10% of the data. Consequently we obtain as well for FFT as for DFT 50 coefficients per frame to capture the mouth shape.

4. Results

In this section the different features will be evaluated via statistical tests and classification experiments in respect to how well they are able to distinguish the two focus conditions. Therefore, in the following the prominence of a word is evaluated only as a binary feature to distinguish these two conditions.

4.1. Statistical Analysis

The first statistical analysis is a paired t-test. It evaluates if for identical words the features in both conditions differ. As can be seen from table 2 this is the case for duration, energy and f0 for all speakers. Spectral emphasis is not significantly different for speaker B. He had significant portions with a hoarse voice where the pitch tracking did not work properly and subsequently also the spectral emphasis feature was impaired. For the DCT and FFT features the second largest coefficient was tested. As well for DCT as FFT this is the coefficient with the lowest vertical frequency which is related to jaw and lip opening. The DCT features are significantly different for all speakers and the FFT for all except for speaker C. Speaker C was turning frequently on the swivel chair during the experiment which could explain why the FFT features are not statistically different. The derivatives of the DCT and FFT show mixed results with no clear tendency. The relative nose y position is not statistically different for any speaker in the two focus conditions. However the nose acceleration for speaker A is.

On one hand these results show that the acoustic features known from the literature to convey word prominence are different for the two conditions and hence the subjects produced the word they corrected with higher prominence. In particular for these words the duration was longer as well as energy, f0 and spectral emphasis higher. On the other hand the results also show that the DCT and FFT coefficients are also able to capture this information.

The paired t-test presented so far requires the knowledge of the word identity as it compares the same words in the two conditions. In order to automatically detect prominent words features which are independent of the word identity are better suited. This can be assessed by an unpaired t-test. It does not compare the features directly in both conditions but the distribution of the features. When looking at the corresponding results in the lower part of table 2 we see that most of the features which were significantly different for the paired t-test also are for the unpaired one. The most notable difference are duration for Speaker A and the DCT coefficient for speaker B and C.

From these results it can be expected that the features are able to discriminate prominent from non-prominent words independent of word identity.

4.2. Detection accuracies

In a step towards a better quantification of the results and the development of a system for automatic audio-visual word prominence detection we trained a Support Vector Machine (SVM) with a Radial Basis Function Kernel using LibSVM [23] to discriminate the two focus conditions. For each feature combination a grid search for the parameters γ and C using the whole dataset was performed. Prior to the grid search the data was normalized to the range $[-1 \dots 1]$. With the found optimal parameters an SVM was trained on 75% of the data and tested on the remaining 25%. Hereby a 30 fold cross validation was used and the dataset was always split such that an identical number of elements was taken from both classes. This process was performed individually for each speaker.

In Fig. 2 the results are shown when only acoustic features are used. The classification based on energy or duration yields accuracies of $60-65\%$ (keep in mind that chance level is 50%). For speaker A the results for spectral emphasis (SE) are similar but as we already discussed in Sec. 4.1 due to frequent segments with hoarse voice they are notably inferior for speaker B and C. For the same reason f0 is a very strong feature for speaker A but

Paired t-test

	dur.	en	f0	SE	y	$y+\Delta$	$y+\Delta\Delta$	DCT	DCT+Δ	DCT+$\Delta\Delta$	FFT	FFT+Δ	FFT+$\Delta\Delta$
A	0.8	< 0.1	< 0.1	< 0.1	6.5	14.1	4.6	< 0.1	< 0.1	7.3	< 0.1	< 0.1	< 0.1
B	< 0.1	< 0.1	< 0.1	11.9	94.0	86.3	38.6	4.7	32.3	58.8	0.7	< 0.1	< 0.1
C	< 0.1	< 0.1	< 0.1	0.2	12.1	19.1	5.2	0.1	0.2	10.4	54.2	9.1	62.9

Unpaired t-test

A	9.6	< 0.1	< 0.1	< 0.1	12.4	17.9	6.7	0.4	3.2	30.3	0.3	0.3	1.1
B	< 0.1	< 0.1	< 0.1	17.2	94.2	87.2	46.4	9.8	44.3	66.7	1.6	0.5	2.4
C	< 0.1	< 0.1	< 0.1	1.1	11.7	25.9	8.4	5.2	2.8	17.2	72.5	27.1	70.1

Table 2: Probabilities in % (p-values$\times$100) that the respective acoustic or visual feature in the two focus conditions is identical calculated either via a paired (top) or an unpaired (bottom) t-test. Features which were not found to differ significantly at a level of 5% are set with grey background. For all transformations the second feature dimension was used. See table 1 for the meaning of the feature name abbreviations.

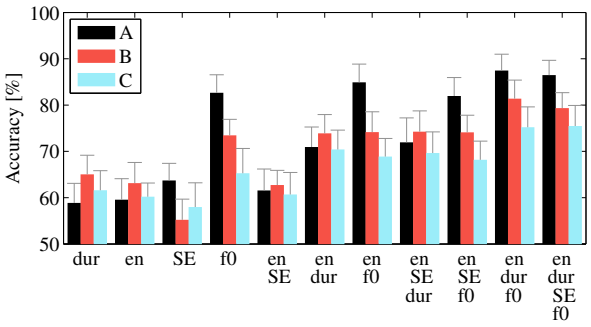

Figure 2: Discrimination accuracies for the different acoustic features and their combination for each speaker. The horizontal lines indicate the standard deviation of the 30 fold cross validation. See table 1 for an explanation of the abbreviations.

less for the other speakers.

When combining different acoustic features we observe that a combination of energy and spectral emphasis does not yield notable improvements. Yet a combination of energy and duration improves the individual results and is for speaker B as good as f0 and for speaker C even better. The combinations with f0 are mainly dominated by this feature. In these cases only duration yields notable improvements. Adding spectral emphasis has in all cases only a minor effect.

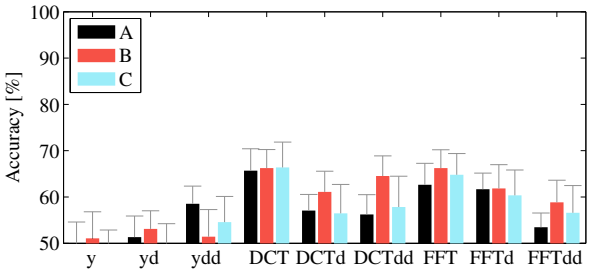

Figure 3: Discrimination accuracies for the different visual features for each speaker. The horizontal lines indicate the standard deviation of the 30 fold cross validation. See table 1 for an explanation of the abbreviations.

The results for the visual features are displayed in Fig. 3. As expected from the analysis in Sec. 4.1 the nose y position and its first and second derivative do not yield good discrimination accuracies. However, DCT and FFT coefficients yield similar or better performance than energy, duration or spectral emphasis. Yet their derivatives show notably inferior performance.

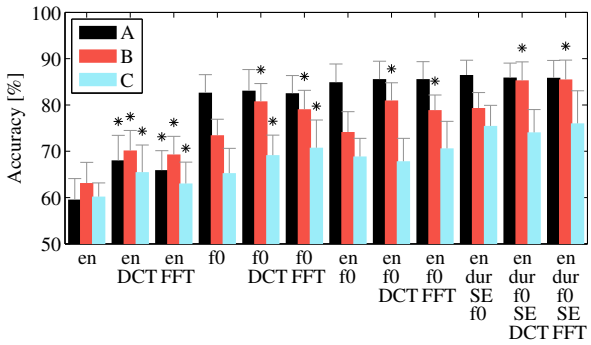

Figure 4: Discrimination accuracies for different acoustic and audio-visual feature combinations for each speaker. The horizontal lines indicate the standard deviation of the 30 fold cross validation. The asterisk indicates results where the combination of acoustic and visual features was found, based on an unpaired t-test, statistically different from the corresponding acoustic only results. See table 1 for an explanation of the abbreviations.

Finally, in Fig. 4 results for acoustic only and a combination of acoustic and visual features are given. When combining energy with DCT or FFT a significant improvement can be seen compared to either feature alone. A combination of f0 and DCT or FFT is mainly dominated by f0 for speaker A but speaker C and in particular speaker B, where f0 was a weaker cue, benefit largely from the additional visual information. The trend is similar when also energy is included. When all acoustic features are combined with DCT or FFT we obtain the best overall results where accuracies for speaker B are almost as good as for speaker A despite the weaker f0 cue.

5. Conclusion

The statistical analysis as well as the expected acoustic feature variation showed that the subjects produced the words in the narrow focus condition with higher prominence. The classification experiments demonstrated that it is possible to distinguish the two focus conditions. A combination of acoustic and visual cues based on DCT or FFT gave the best results with accuracies of $74 - 86\%$, depending on the speaker. We also investigated head movements via the nose x and y position, including speed and acceleration. For the nose features we saw no significant differences between the two focus conditions and also only low detection accuracies when taken alone and no gain when combined with other features. In other studies the head movements

were sometimes found to be a consistent cue for prosodic focus [15] and sometimes not [14].

The acoustic features duration and energy, also in combination, showed good and consistent performance. For f0 the differences between the three speakers were quite large. For speaker A, the female speaker, f0 was a very strong cue and adding other features to it only slightly improved the results. In particular the additional visual information led only to very small improvements. As already discussed f0 was a notably weaker cue for speaker B and more so for speaker C. Due to this weaker performance of f0 the integration of visual information yielded larger improvements for these speakers. When comparing for speaker B the accuracies for f0 alone and in combination with the DCT they increase from 73% to 81%. This increase is statistically significant. For speaker C they increase for f0 alone and in combination with the DCT from 65% to 69%, a statistically not significant increase. This lower overall detection performance for speaker C can in addition to the hoarse voice also be attributed to its less consistent prominence production. In several occasions he already made words in the original utterance highly prominent.

In general the DCT yielded slightly better results than the FFT but not for speaker C. Overall our subjects tilted their heads quite frequently during the interaction. The current visual feature extraction does not take care of this. This impairs as well the DCT as the FFT features as they are not rotation invariant. Significant improvements of the visual features can be expected once this tilt is corrected.

For both DCT and FFT the second largest coefficient, with the largest being the mean energy, was the one capturing the lowest vertical frequency. This is closely related to the jaw and lip opening. The results in Sec. 4.1 showed that this coefficient differs significantly for the two focus conditions which is in agreement with previous findings that narrow focus yields to hyper articulation for these articulators [14].

Overall we saw via this experiment that it is possible to extract visual prosodic cues without additional markers and that these cues yield by themselves and in combination with acoustic cues good accuracies for the discrimination of prominent from non-prominent words without knowledge of the word identity. This constitutes to the authors best knowledge the first result showing that classification of prominent and non-prominent words can be improved by including visual cues. More speakers have to be tested in the future to further substantiate these results.

6. Acknowledgments

I want to thank Petra Wagner, Britta Wrede and Heiko Wersing for fruitful discussions. Furthermore, I am 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. Special thanks go to our subjects for their patience and effort.

7. References

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

[2] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "Verbmobil: The use of prosody in the linguistic components of a speech understanding system," *IEEE Trans. Speech and Audio Process.*, vol. 8, no. 5, pp. 519–532, 2000.

[3] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C.V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[4] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech Communication*, vol. 43, no. 1-2, pp. 155–175, 2004.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] M. Wolters and P. Wagner, "Focus perception and prominence," in *Computers, Linguistics, and Phonetics between Language and Speech. Proc. 4th Conf. on Natural Language Process.*, 1998, vol. 1, pp. 227–236.

[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, pp. 2421, 2006.

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

[20] M. Heckmann, C. Gläser, M. Vaz, T. Rodemann, F. Joublin, and C. Goerick, "Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Systems (IROS)*, Nice, 2008.

[21] M. Heldner, "On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in swedish," *Journal of Phonetics*, vol. 31, no. 1, pp. 39–62, 2003.

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

[23] 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`.