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

DIFFERENCES BETWEEN SPEAKERS IN AUDIO-VISUAL CLASSIFICATION OF WORD PROMINENCE

Martin Ernst Heckmann

2013

Preprint:

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

DIFFERENCES BETWEEN SPEAKERS IN AUDIO-VISUAL CLASSIFICATION OF WORD PROMINENCE

Martin Heckmann
Honda Research Institute Europe GmbH
D-63073 Offenbach/Main, Germany
martin.heckmann@honda-ri.de

Abstract: We show how the audio-visual discrimination performance of prominent from non-prominent words based on an SVM classifier varies from speaker to speaker. We collected data in an experiment where users were interacting via speech in a small game, designed as a Wizard-of-Oz experiment, with a computer. Following misunderstandings of one single word of the system, users were instructed to correct this word using prosodic cues only. Hence we obtain a dataset which contains the same word with normal and with high prominence. Overall we recorded 8 speakers. The analysis shows that there is a large variation from speaker to speaker in respect to which feature can successfully be used to discriminate prominent from non-prominent words depending on the prominence signaling strategy applied by the speaker. In particular for speakers who mainly use duration to signal prominence we see an increase in performance from combining acoustic and visual information. The audio-visual classification accuracies we obtain vary from $66\% - 91\%$ correct from the most difficult to the easiest speaker.

1 Introduction

Current spoken dialog systems don't evaluate the prosodic characteristics of speech even though it is well known that prosodic cues play a very important role in human communication [21]. Nevertheless, quite a few research systems included such prosodic cues in a human-machine dialog [19, 22, 15]. In general the inclusion of prosodic cues is quite difficult as they show not only a large variability from speaker to speaker but are also difficult to extract from the speech signal. The inclusion of visual information might be a route to alleviate these problems. Information on the movements of the speaker's mouth and face notably improves the accuracies of automatic speech recognition, particular in difficult situations [20, 11, 16, 24]. Humans are also able to use such visual information to extract prosodic cues [9, 18, 2, 23, 1]. Studies quantifying these visual prosodic cues have shown that they are mainly manifested in larger jaw opening, lip spreading and protrusion and to some extend to head movements [8, 7].

In [17] it was shown that speakers use prosodic cues to highlight corrections in a dialog with a machine and that these can be detected using prosodic cues. We extended this idea in [10] to the audio-visual discrimination of prominent from non-prominent words. In particular we showed that the performance can be improved by visual features extracted from the speaker's face without the use of additional visual markers. As visual features we used image transformations calculated on the mouth region of the speaker. For this paper we extended the dataset and investigate how the discrimination of prominent from non-prominent words varies from speaker to speaker.

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. In the last section we will discuss the results.

2 Dataset

For the recording of the data the subjects interacted via speech in a Wizard of Oz experiment with a computer in a small game where they would move tiles to uncover a cartoon. With this playful setting we expected to obtain more natural speech, in particular regarding the prosody. This game yielded utterances of the form 'put green in B one'. 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 [3]. The subjects were told to repeat in these cases the phrase as they 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.

In total 8 subjects, 4 females and 4 males, three speaking British English as their sole native language, three being bilingual British English/German, one speaking American English as her sole native language and one being bilingual American 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 [6] a speech recognition system trained on that corpus could be used to perform a forced alignment on the acquired data. For the alignment HTK and a combination of RASTA-PLP and spectro-temporal HIST features [12] was used as this gave upon visual inspection better results then either of the feature sets alone or MFCC features. In particular, we first performed a speaker adaptation with a Maximum Likelihood Linear Regression (MLLR) step followed by a Maxium A-Posteriori (MAP) step, both using HTK [25].

For further processing those turns where the original utterance and a correction were available were selected. This yielded overall 1300 turn pairs (original utterance + correction), i. e. on average ≈ 160 turn pairs per speaker). 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. An analysis of acoustic features related to word prominence in [10] showed that the words in the narrow focus condition were notably more prominent than in the broad focus condition .

3 Features

In the following experiments the features described in Table 1 which have previously been proposed to capture word prominence were used. From these features (except for duration) 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.

For the visual modality the openCV library [4] 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 and hence also on the current position of the mouth region. 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

Acoustic

dur	duration of the word
en	energy relative to the mean of the utterance
f0	mean fundamental frequency (extracted according to [14, 13])

Visual

y	nose y position relative to the mean of the utterance
d, dd	first and second derivative

Table 1 - A description of the different features.

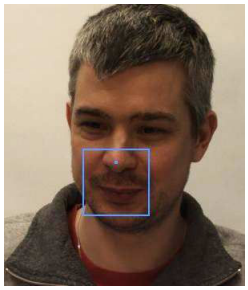


Figure 1 - Image from recording after cropping to face region, nose detection, downsampling and highlighting of the mouth region.

was kept identical. After downsampling by a factor of 2 this yields an image of 80×80 pixels of the mouth region (compare Figure 1). On these images either a two-dimensional Fast Fourier Transform (FFT) or Discrete Cosine Transform (DCT) was calculated. In case of the FFT and DCT out of the 6400 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 6400 coefficients on a randomly selected subset of 10% of the data. As FFT coefficients are complex we only used their magnitude in all steps. Consequently we obtain for FFT and DCT 50 coefficients per frame to capture the mouth shape. 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.

4 Results

To discriminate prominent from non-prominent words a Support Vector Machine (SVM) with a Radial Basis Function Kernel was trained using LibSVM [5]. 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, was performed using the whole dataset. 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 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.

When looking on the results in Figure 2 we can see a large variation in performance for the different features and also for the different speakers. Overall duration and fundamental frequency perform identical with a 65% correct rate and energy inferior with 59% correct. Yet the variation from speaker to speaker for all features is very large. For duration the speaker yielding the best results is speaker F with 76% correct. The two speaker with the worst results, speaker A and C, obtain only 53% and 57% correct, respectively. However, for fundamental frequency speaker C is actually the one yielding with 83% correct the best results. Also speaker A is with 69% correct above average. When looking also on the results for the energy feature one can see that speaker A seems to use mainly fundamental frequency to signal prominence. The results suggest that most speakers use either duration or fundamental frequency to signal prominence and not both at the time.

Combining the different acoustic features improves the performance. The combination of energy and duration yields 70% correct and adding also fundamental frequency increases perfor-

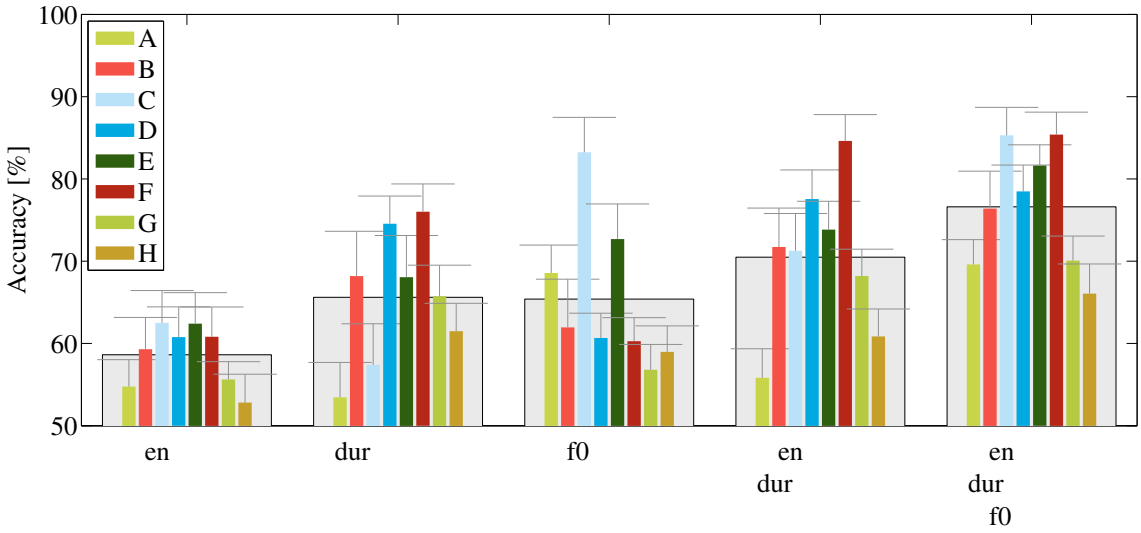


Figure 2 - Discrimination accuracies for different acoustic feature combinations. The grey bars in the background visualizes the average over all speakers for a given feature or feature combination. The short horizontal lines indicate the standard deviation of the 30 fold cross validation. See Table 1 for an explanation of the abbreviations.

mance to 77% correct. For the combination of all three features the inter speaker variation is reduced as in this case the SVM is able to yield good results if they use duration or fundamental frequency.

In Figure 3 the results using the different visual features are displayed. As we can see on average FFT and DCT perform similar, yet with 66% correct DCT is somewhat better than FFT with 64% correct. With a range of $56\% - 89\%$ correct for FFT the performance varies a lot from speaker to speaker. When looking on the nose y position we can see that it is for most speakers non-informative. However, with 70% correct speaker F clearly stands out. This speaker also yields the best results for FFT and DCT.

When comparing the results obtained by acoustic and visual features one can see that in particular for speakers which use duration to signal prominence it can also be identified well from the visual channel. Yet it seems that it is not only the duration or a combination of duration and energy which is extracted from the visual channel as for speaker F, G and H the results using the visual channel only are better than those for duration or the combination of energy and duration. The corresponding results are depicted in Table 2.

Table 2 - Classification rates in %.

speaker	duration	energy+duration	DCT
F	76	85	86
G	66	68	72
H	62	61	65

Finally in Figure 4 the results we obtain when we combine the acoustic and the visual features are depicted. Here we can see that when combining either energy and duration with FFT, respectively DCT, or the combination of all acoustic features with FFT, respectively DCT we can see for some speakers a significant increase in performance. As we can expect from the previous results the results for speaker F and G improve from the combination of acoustic and visual information (from 85% to 91% and from 70% to 77% correct for the combination of all

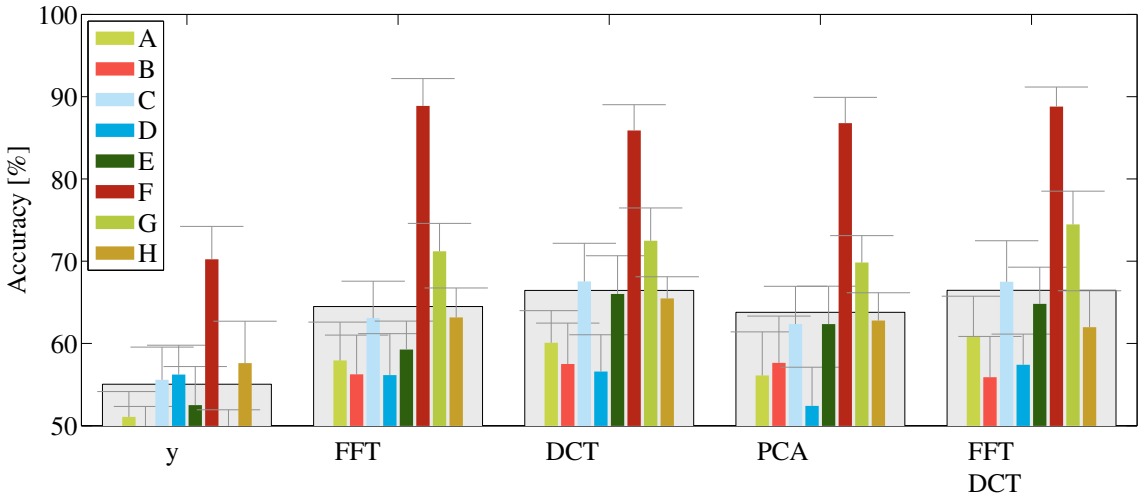


Figure 3 - Discrimination accuracies for different visual feature combinations. The grey bars in the background visualizes the average over all speakers for a given feature or feature combination. The short horizontal lines indicate the standard deviation of the 30 fold cross validation. See Table 1 for an explanation of the abbreviations.

acoustic features either without or with DCT features). Yet also the classification performance for speaker A who, as we saw, mainly uses fundamental frequency, increases from the addition of the visual channel (from 70% to 73% correct).

5 Conclusion

We collected via a small Wizard-of-Oz game with a computer data where subjects were uttering words with normal and high prominence. During the game we made audio and video recordings. We then extracted acoustic and visual features and trained an SVM classifier to discriminate the normal from the highly prominent words. The results showed that the different speakers used different strategies to indicate prominence, i. e. chiefly via duration or fundamental frequency. One speaker also very consistently used head movements. We also saw that those speakers where the visual discrimination results were best showed in the acoustic channel a preference for duration. This is intuitive as changes during phonation, i. e. fundamental frequency, are not visible. However, we also saw that for some speakers the visual channel provides more information than the acoustic channel. In addition to the longer opening time of the mouth due to an increase of segment duration this is most likely information on hyper-articulation, e. g. wider opening of the mouth or a wider spreading of the lips [8]. In particular for these speakers the inclusion of the visual channel increases the discrimination performance compared to the audio channel alone. Next steps include the evaluation of the whole utterance instead of only the prominent word to include e. g. hypo-articulation effects of the following words [8].

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

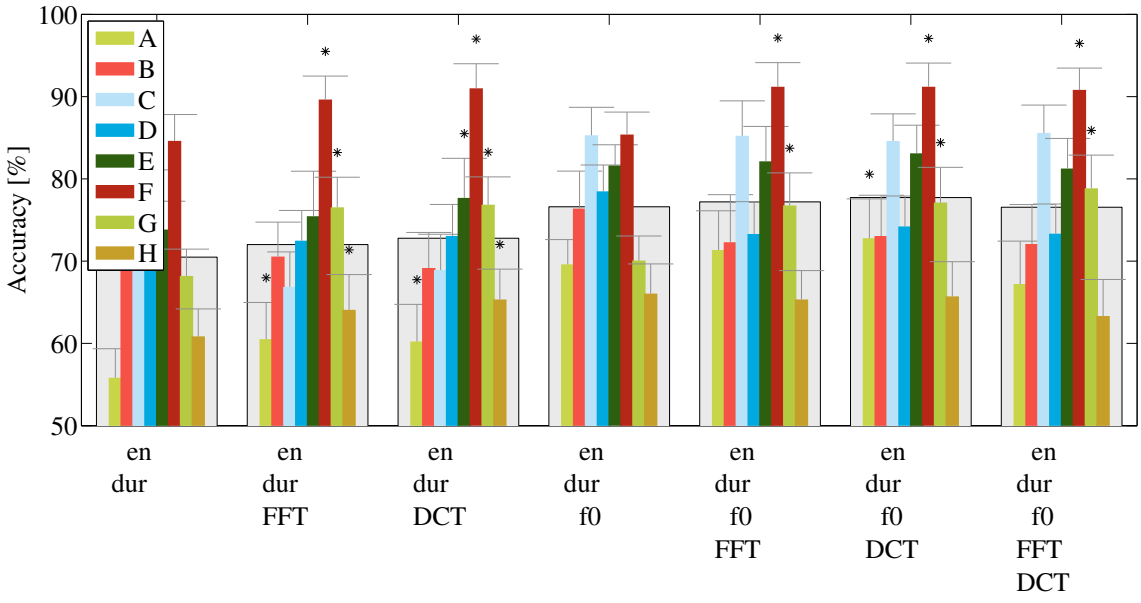


Figure 4 - Discrimination accuracies for different acoustic and visual feature combinations. The grey bars in the background visualizes the average over all speakers for a given feature or feature combination. The short horizontal lines indicate the standard deviation of the 30 fold cross validation. The asterisk indicates audio-visual results which are statistically significantly better ($\alpha = 0.05$) than the corresponding audio only results. See Table 1 for an explanation of the abbreviations.

tuning the SVMs and Merikan Koyun for help in the data preparation. Special thanks go to my subjects for their patience and effort.

References

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

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

[3] A. Black, P. Taylor, and R. Caley. The festival speech synthesis system. Technical report, 1998.

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

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

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

[7] 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.

[8] 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.

[9] H. Graf, E. Cosatto, V. Strom, and F. Huang. Visual prosody: Facial movements accompanying speech. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 396–401. IEEE, 2002.

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

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

[12] M. Heckmann, X. Domont, F. Joublin, and C. Goerick. A hierarchical framework for spectro-temporal feature extraction. *Speech Communication*, 53(5):736 – 752, 2011. Perceptual and Statistical Audition.

[13] 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.

[14] M. Heckmann, F. Joublin, and C. Goerick. Combining rate and place information for robust pitch extraction. In *Proc. INTERSPEECH*, pages 2765–2768, Antwerp, 2007.

[15] J. Hirschberg, D. Litman, and M. Swerts. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43(1-2):155–175, 2004.

[16] 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.

[17] G. Levow. Identifying local corrections in human-computer dialogue. In *Eighth International Conference on Spoken Language Processing*, 2004.

[18] K. Munhall, J. Jones, D. Callan, T. Kuratate, and E. Vatikiotis-Bateson. Visual prosody and speech intelligibility. *Psychological Science*, 15(2):133, 2004.

[19] 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.*, 8(5):519–532, 2000.

[20] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE*, 91(9):1306–1326, 2003.

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

[22] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[23] M. Swerts and E. Krahmer. Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, 36(2):219–238, 2008.

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

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