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

Visual Contribution to Word Prominence Detection in a Playful Interaction Setting

Martin Ernst Heckmann

2013

Preprint:

This is an accepted article published in Natural Interaction with Robots, Knowbots and Smartphones - Putting Spoken Dialog Systems into Practice,. The final authenticated version is available online at: https://doi.org/[DOI not available]

Visual Contribution to Word Prominence Detection in a Playful Interaction Setting

Martin Heckmann

Abstract This paper investigates how prominent words can be distinguished from non-prominent ones in a setting where a user was interacting in a small game, designed as a Wizard of Oz experiment, with a computer. Misunderstandings of the system were triggered and the user was asked to correct them naturally, i. e. using prosodic cues. Consequently, the corrected word is expected to be highly prominent. Audio-visual recordings with a distant microphone and without visual markers were made. As acoustic features relative energy, duration and fundamental frequency were calculated. From the visual channel rigid head movements and image transformation based features from the mouth region were extracted. Different feature combinations are evaluated regarding their power to discriminate the prominent from the non-prominent words using a Support Vector Machine (SVM). Depending on the features accuracies of approx. $70-80\%$ are achieved. Thereby the visual features are in particular beneficial when the acoustic features are weaker.

1 Introduction

Current commercial spoken dialog systems are insensitive to prosodic characteristics of speech even though it is well known that prosodic cues play a very important role in human communication (Shriberg, 2005). Nevertheless, quite a few research systems included such prosodic cues in a human-machine dialog (Nöth et al, 2000; Stolcke et al, 2000; Hirschberg et al, 2004). 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 latter problem could possibly be reduced by including visual information. Information on the movements of the speaker's mouth and face notably improves the accuracies of automatic speech recognition, particular in difficult situations (Potamianos et al,

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

2003; Heckmann et al, 2002; Kolossa et al, 2009; Yoshida et al, 2009). Humans are also able to use such visual information to extract prosodic cues (Graf et al, 2002; Munhall et al, 2004; Beskow et al, 2006; Swerts and Krahmer, 2008; Al Moubayed and Beskow, 2009). 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 (Dohen et al, 2006; Cvejic et al, 2010).

In Heckmann (2012) we showed that the discrimination of prominent from non-prominent words 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 improved the feature extraction and investigate how rigid head movements and dynamic features can support this discrimination.

In the next section an overview on the recording of the data will be given. After that Sec. 3 describes the different features extracted from the acoustic and visual channel. Following this Sec. 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 (Black et al, 1998). 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 3 subjects, one female and two males, 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 (Cooke et al, 2006) 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 (Heckmann et al, 2011) 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 (Young et al, 1995).

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. An analysis of acoustic features related to word prominence in Heckmann (2012) 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.

Acoustic	
dur	duration of the word
en	energy relative to the mean of the utterance
f0	mean fundamental frequency (extracted according to Heckmann et al (2007, 2008))
Visual	
y	nose *y* position relative to the mean of the utterance
d, dd	first and second derivative

Table 1 Description of the different features.

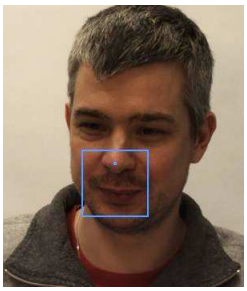


Fig. 1 Image from recording after cropping to face region, nose detection, downsampling and highlighting of the mouth region.

For the visual modality the openCV library (Bradski, 2000) 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 was kept identical. After downsampling by a factor of 2 this yields an image of 80×80 pixels of the mouth region (compare Fig. 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 (Chang and Lin, 2011). 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 $[-1\ldots1]$. 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. In the following all results are averaged over all speakers.

When taking only the acoustic cues into account we see that the relative energy achieves a performance of 63%, duration of 59% and f0, a notably stronger cue, 70% correct. The combination of all acoustic cues achieves 80% correct.

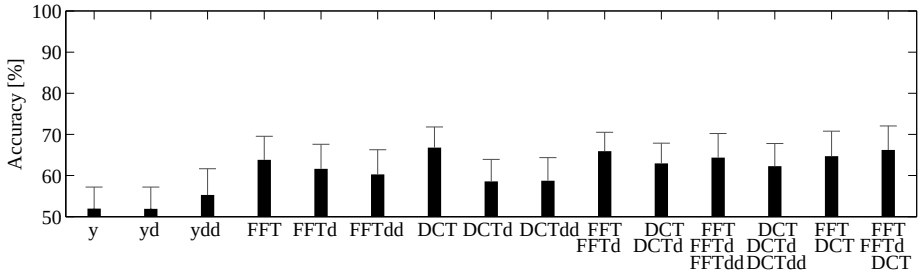


Fig. 2 Discrimination accuracies for different visual feature combinations. The horizontal lines indicate the standard deviation of the 30 fold cross validation. See table 1 for an explanation of the abbreviations.

If we now look at the results for the visual cues in Fig. 2 we see that FFT achieves 64% and DCT 67% correct. Adding the first derivative to the FFT improves its performance to 66% correct, almost the level of the DCT. Adding the second derivative has a negative effect. For the DCT all derivatives have a negative effect. Combining FFT and DCT features does not yield to superior performance than DCT alone.

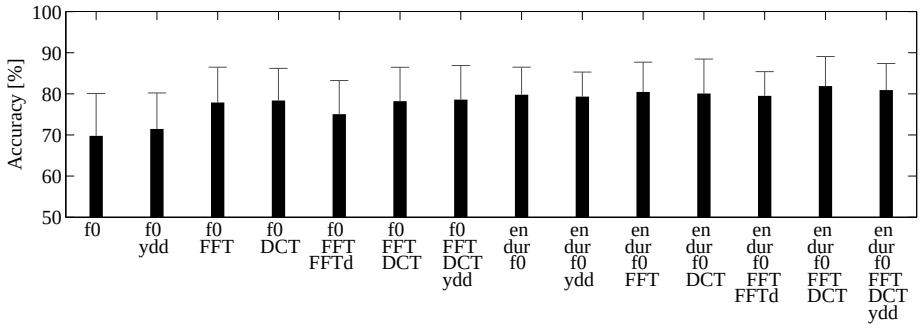


Fig. 3 Discrimination accuracies for different acoustic and visual feature combinations. The horizontal lines indicate the standard deviation of the 30 fold cross validation. See table 1 for an explanation of the abbreviations.

Regarding the nose *y* movement, only the second derivative, i. e. the acceleration of the nose, yields accuracies notably above chance level.

In Fig. 3 a combination of the f0 feature, the strongest acoustic feature, and different visual features is depicted. Adding the nose *y* acceleration to the f0 feature leads to a small improvement. When combining either FFT, DCT or both with f0 the improvement is from 70% to 78%. Including either derivatives of FFT or DCT or the nose *y* acceleration does not further improve the results.

Finally Fig. 3 also shows a combination of all acoustic features and different visual features. The performance of the acoustic features alone is 80% correct. Adding either FFT or DCT alone does not change this. However, when both are added 82% correct are obtained. When adding also the nose *y* acceleration it is reduced to 81% correct.

5 Conclusion

We set up a Wizard of Oz experiment where subjects were interacting in a small game with a computer. The experiment was designed to elicit prominent and non-prominent versions of a word. The results showed that these two classes can be discriminated based on acoustic and visual features with accuracies of $\approx 80\%$. Thereby the nose position seems to contribute some information but much less than the other features and in several cases adding it reduces performance. The same can be said about the derivatives of the FFT and DCT features. When combining the fundamental frequency, in our case the strongest acoustic feature, with visual features a notable improvement can be seen. Yet when all acoustic features are used the further gain from visual features is small. Amongst the acoustic features are duration and relative energy. In particular the former could only reliably extracted due to the forced alignment used. The latter is vulnerable to varying distances to the microphone. Hence these would be available with much less precision in a real world system. In these cases the additional visual features can be very beneficial. Therefore, we think that adding visual features renders the extraction of prosodic cues

more reliable and hence can be helpful to develop more natural spoken dialog systems which can deal meaningfully with prosodic variations. Next steps include the extension of the approach to more speakers and the evaluation of the whole utterance instead of only the prominent word to include e. g. hypo-articulation effects of the following words (Dohen et al, 2006).

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 my subjects for their patience and effort.

References

Al Moubayed S, Beskow J (2009) Effects of visual prominence cues on speech intelligibility. In: Proc. Int. Conf. Auditory Visual Speech Process. (AVSP), ISCA, vol 9, p 16

Beskow J, Granström B, House D (2006) Visual correlates to prominence in several expressive modes. In: Proc. INTERSPEECH, ISCA, pp 1272–1275

Black A, Taylor P, Caley R (1998) The festival speech synthesis system. Tech. rep.

Bradski G (2000) The OpenCV Library. Dr Dobb's Journal of Software Tools

Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2:27:1–27:27, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

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

Cvejic E, Kim J, Davis C, Gibert G (2010) Prosody for the eyes: Quantifying visual prosody using guided principal component analysis. In: Proc. INTERSPEECH, ISCA

Dohen M, Lœvenbruck H, Harold H, et al (2006) Visual correlates of prosodic contrastive focus in french: Description and inter-speaker variability. In: Speech Prosody, Dresden, Germany

Graf H, Cosatto E, Strom V, Huang F (2002) Visual prosody: Facial movements accompanying speech. In: Int. Conf. on Automatic Face and Gesture Recognition, IEEE, pp 396–401

Heckmann M (2012) Audio-visual evaluation and detection of word prominence in a human-machine interaction scenario. In: Proc. INTERSPEECH, ISCA, Portland, OR

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

Heckmann M, Joublin F, Goerick C (2007) Combining rate and place information for robust pitch extraction. In: Proc. INTERSPEECH, Antwerp, pp 2765–2768

Heckmann M, Gläser C, Vaz M, Rodemann T, Joublin F, Goerick C (2008) 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

Heckmann M, Domont X, Joublin F, Goerick C (2011) A hierarchical framework for spectro-temporal feature extraction. Speech Communication 53(5):736 – 752, DOI DOI: 10.1016/j.specom.2010.08.006, perceptual and Statistical Audition

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

Kolossa D, Zeiler S, Vorwerk A, Orglmeister R (2009) Audiovisual speech recognition with missing or unreliable data. In: Proc.Int. Conf.Auditory Visual Speech Process. (AVSP)

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

Nöth E, Batliner A, Kießling A, Kompe R, Niemann H (2000) Verbmobil: The use of prosody in the linguistic components of a speech understanding system. IEEE Trans Speech and Audio Process 8(5):519–532

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

Shriberg E (2005) Spontaneous speech: How people really talk and why engineers should care. In: Proc. EUROSPEECH, ISCA

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

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

Yoshida T, Nakadai K, Okuno H (2009) Automatic speech recognition improved by two-layered audio-visual integration for robot audition. In: Proc. 9th IEEE-RAS Int. Conf. on Humanoid Robots, IEEE, pp 604–609

Young S, Odell J, Ollason D, Valtchev V, Woodland P (1995) The HTK Book. Cambridge University, Cambridge, United Kingdom