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

Prosodic, Spectral and Visual Features for the Discrimination of Prominent and Non-prominent words

Martin Ernst Heckmann

2013

Preprint:

This is an accepted article published in 40. Jahrestagung für Akustik (DAGA). The final authenticated version is available online at: https://doi.org/[DOI not available]

Prosodic, Spectral and Visual Features for the Discrimination of Prominent and Non-prominent Words

Martin Heckmann

Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Email: martin.heckmann@honda-ri.de

Introduction

Despite its very high relevance for human communication current spoken dialog systems usually ignore the prosodic variations in the speech signal [1, 2, 3]. In [4] 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 [5] to the audio-visual discrimination of prominent from non-prominent words. Visual features have been shown to play an important role for human perception of prosody [6]. In this paper we have a closer look on the information contained in the different features from the acoustic and visual channel. In particular, we investigate the contribution of the visual features, i. e. nose movement and mouth appearance, and spectral features.

Database

We recorded a database where subjects interacted via speech in a Wizard of Oz experiment with a computer in a small game, yielding utterances of the form 'place green in B one'. Following a misunderstanding by the system, induced by the operator, subjects were asked to correct it using only prosodic cues (see [7] for details). Following a manual annotation we selected from the 16 speakers 8 speakers (2 female, 6 male) which were annotated most consistently. 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. In order to have temporal alignments of the data we used a speech recognition system trained on a similar task and performed a forced alignment using the features presented in [8].

For further processing those turns where the original utterance and a correction were available were selected. Overall we have 1263 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 on a subset of the data in [5] showed that the words in the narrow focus condition were notably more prominent than in the broad focus condition .

Features

We extracted different features from the acoustic and visual channel to capture the prosodic variations.

Acoustic features

From the acoustic channel we extracted the logarithmic short term energy e, the duration of the word and the gaps before and after the word as determined from the forced alignment D, the fundamental frequency f_0 (following [9]), spectral emphasis SE, i. e. the difference between the overall intensity and the intensity in a dynamically low-pass-filtered signal with a cut-off frequency of $1.5 f_0$ [10], formant frequencies F [11] and RASTA-PLP features (RASTA) [12] which are originally designed to capture the phonetic content of the signal. Where appropriate we normalized the features by their utterance mean and extracted the max, min, mean, var and spread (max-min) of the feature and its first and second derivatives.

Visual features

To extract features from the visual channel we used the openCV library [13]. By its help we detected the nose in each image. Based on this we developed a tracking algorithm which yields the nose position over time. Starting from the nose we can determine the mouth region in the image via a fixed speaker independent offset from the nose. On each subsampled mouth image of size 100×100 pixels we calculated a two-dimensional Discrete Cosine Transform (DCT). Out of the 10000 coefficients per image we selected the 50 with the lowest spatial frequencies. We also normalized the visual features by their utterance mean and extracted the max, min, mean, var and spread (max-min) of the feature and its first and second derivatives.

Results

To evaluate the different features we used an SVM in a 30 fold cross-validation (using 75% for training and 25% for testing [7]) to discriminate prominet from non-prominent words. As can be seen in the upper part of Fig. 1 fundamental frequency and energy are with 74.2% and 74.5% correct the strongest features. RASTA features perform very similar (74.2%). Also the visual features (nose 68.8% and DCT 69.7%) are quite strong features, slightly better than duration (68.6%) and formants (67.2%). Spectral emphasis is rather weak (57.4%).

When looking on the combination of the features in the lower part of Fig. 1 we see that combining the features well known to capture prosody, i. e. energy, duration, spectral emphasis and fundamental frequency, we obtain 81.6% correct. Adding the formant features does not improve this result in general but for one speaker. Adding

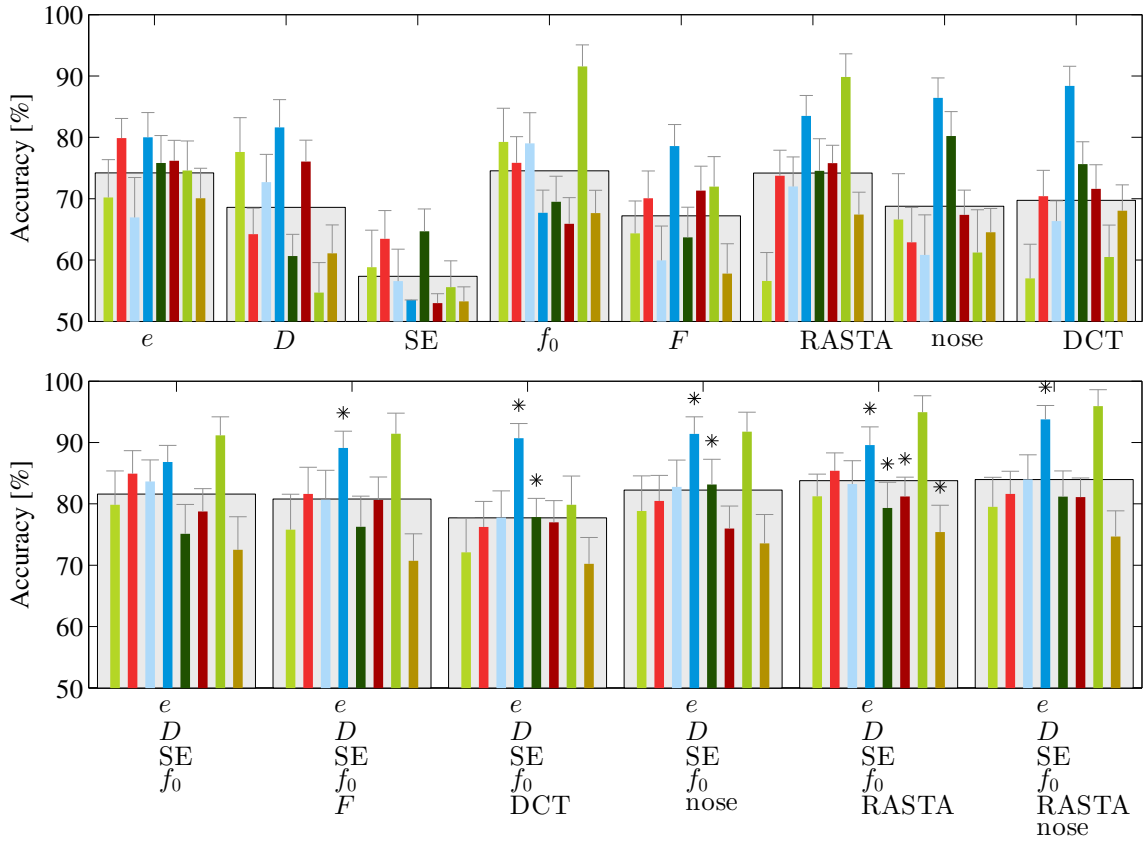


Figure 1: Discrimination accuracies between prominent and non-prominent words. The thin colored lines show the results for the individual speakers and the grey thick bar in the background the average over all speakers. The asterisk indicates statistically significant improvements at $\alpha = 0.05$.

the DCT features has on average a negative effect. When also using the nose features the performance improves slightly. Yet for the individual speakers the additional visual information can have a very strong positive effect (e. g. results for the fourth speaker increase from 86.8% to 90.7% and 91.4%, respectively).

Discussion

We could show that the visual channel carries a lot of information on the prosodic content and that discrimination accuracies are for all speaker for nose movement and mouth DCT well above change level. Overall the contribution of the visual channel is very speaker dependent. Many of the features are strongly influenced by the actual word (e. g. duration). For RASTA features we see a very high risk of a specialization of the features to the word and the production of the word (i. e. prominent or non-prominent). We consider this risk as particulary high as the RASTA features have 39 dimensions yet there are only 34 words in our dataset.

References

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

[2] 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.

[3] 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.

[4] G.A. Levow, "Identifying local corrections in human-computer dialogue," in *Eighth Int. Conf. on Spoken Lang. Proc. (ICSLP)*, 2004.

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

[6] 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.

[7] M. Heckmann, "Inter-speaker variability in audio-visual classification of word prominence," in *Proc. INTERSPEECH*, Lyon, France, 2013.

[8] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A closer look on hierarchical spectro-temporal features (HIST)," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008.

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

[10] 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.

[11] C. Gläser, M. Heckmann, F. Joublin, and C. Goerick, "Combining auditory preprocessing and bayesian estimation for robust formant tracking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 224–236, 2010.

[12] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans Speech and Audio Proc.*, vol. 2, no. 4, pp. 578–589, 1994.

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