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

Image Transformation based Features for the Visual Discrimination of Prominent and Non-Prominent Words

Martin Ernst Heckmann

2012

Preprint:

This is an accepted article published in Proc. 10th ITG Conference on Speech Communication. The final authenticated version is available online at: https://doi.org/[DOI not available]

Image Transformation based Features for the Visual Discrimination of Prominent and Non-Prominent Words

Martin Heckmann

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

Abstract

This paper investigates how visual information extracted from a speaker's mouth region can be used to discriminate prominent from non-prominent words. The analysis relies on a database where users interacted in a small game with a computer in a Wizard of Oz experiment. Users were instructed to correct recognition errors of the system. This was expected to render the corrected word highly prominent. Audio-visual recordings with a distant microphone and without visual markers were made. As acoustic features relative energy and fundamental frequency were calculated. From the visual channel image transformation based features from the mouth region were extracted. As image transformations FFT, DCT and PCA with a varying number of coefficients are compared in this paper. Thereby the performance of the visual features by themselves or in combination with the acoustic features is investigated. The comparison is based on the classification with a Support Vector Machine (SVM). The results show that all three image transformations yield a performance of approx. 65% in this binary classification task. Furthermore, the information extracted from the visual channel is complementary to the acoustic information. The combination of both modalities significantly improves performance up to approx. 80%.

1 Introduction

It is generally accepted that prosody, the rhythm, stress, and intonation of speech, plays a very important role in human-human communication. Despite several attempts up to date it is rarely used in spoken dialog systems [1–4]. The fact that prosodic cues are notoriously difficult to extract reliably certainly is one of the reasons for that. Not only are they manifested in fragile acoustic cues (e. g. relative energy and duration) but they also show a large inter and intra speaker variability.

The observation by Sumby and Pollack [5] that human speech perception in noise is substantially better when information on the movement of the speaker's face is available stimulated a lot of research to integrate the visual modality in automatic speech recognition systems [6–9]. These systems showed SNR gains in the order of $6-10$ dB from adding the visual channel.

On the other hand it has been shown that the visual channel also conveys relevant information for the perception of prosody [10–14].

This suggests that the inclusion of the visual channel in the extraction of prosodic cues is promising. In [15] we showed that the discrimination of prominent from non-prominent words can be improved by visual features extracted from the speakers face without the use of additional visual markers. As visual features we used image transformations calculated on the mouth region of the speaker.

This paper further investigates the suitability of these image transformations for the extraction of visual prosodic cues and there use for the discrimination of prominent from non-prominent words. In particular it compares the Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT) and Principal Component Analysis (PCA) with varying dimensionalities for the discrimination of prominent from non-prominent words either solely based on visual features or in combination with acoustic features.

In the next section an overview on the data used 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 classification experiments. In the last section we will discuss the results and give an outlook.

2 Dataset

For the recording of the data the subjects interacted in a Wizard of Oz experiment with a computer in a small game where they would move tiles to uncover a cartoon. This yielded utterances of the form "**put green 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 [16]. 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, 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 [17] 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 [18] was used as this gave upon visual inspection better results then either of the features alone or MFCC features.

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

Acoustic	
en	energy relative to the mean of the utterance
f0	mean fundamental frequency (extracted according to [19, 20])
Visual	
y	nose y position relative to the mean of the utterance
Δ, $\Delta\Delta$	first and second derivative

Table 1: Description of the different features.

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 capture word prominence were used. From these features 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 [21] 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

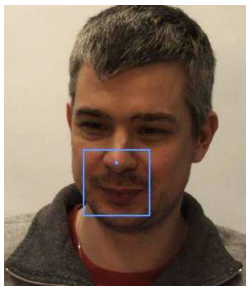

Figure 1: Image from recording after cropping to face region, nose detection, downsampling and highlighting of the mouth region.

a two-dimensional Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), or Principal Component Analysis (PCA) was calculated.

In case of the FFT and DCT out of the 6400 coefficients per image the N 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. The covariance matrix needed to calculate the PCA was also calculated on this 10% subset. For the PCA the N coefficients with the highest eigenvalues were retained. Consequently we obtain for FFT, DCT and PCA N coefficients per frame to capture the mouth shape. The original image of the mouth region and a reconstruction with 50 coefficients using the different image transformations is shown in Fig. 2.

4 Results

To discriminate prominent from non-prominent words a Support Vector Machine (SVM) with a Radial Basis Function Kernel was trained using LibSVM [22]. 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, 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 in which the data set was always split such that an identical number of elements is taken from both classes was run. This process was performed individually for each speaker.

Fig. 3(a) shows the performance of the FFT, DCT, and PCA with varying dimensionality. As can be seen even with only 5 features a detection above chance level (50%) is possible. The best performance for FFT (65%) is achieved with 50 dimensions. The DCT achieves the same performance already with 25 dimensions, the PCA only with $100-250$ dimensions.

In Fig. 3(b) Deltas and Delta-Deltas, i. e. first and second derivative, of the features were calculated and combined with the original features. As can be seen the performance is inferior to the performance without the derivatives.

In Fig. 3(c) the visual features have been combined with the mean relative energy. The results show a significant performance improvement from the combination of both modalities already with 25 or 50 features.

For the results in Fig. 3(d) also the fundamental frequency (f0) information was added to the acoustic channel. As can be seen fundamental frequency by itself is a much stronger cue than the mean relative energy, in particular for speaker 1. Yet the combination of fundamental frequency and the image transformation features further improves performance in most cases. This improvement is statistically significant for DCT with 25 dimensions and FFT and PCA with 50 dimensions.

5 Conclusion

A previous statistical analysis on the same data based on the acoustic features had shown that the subjects indeed produced the words in the narrow focus condition with higher prominence [15]. The current results help to better quantify which of the different image transformations investigated is best suited to the task and how many coefficients of each transformation have to be retained to achieve these results.

The results show that all three image transformations are able to distinguish the two focus conditions well above change level. With all three transformations it is possible to obtain classification accuracies averaged over all speakers of $\approx 65\%$. In general they show a trend for performance increase when more features are used up to an op-

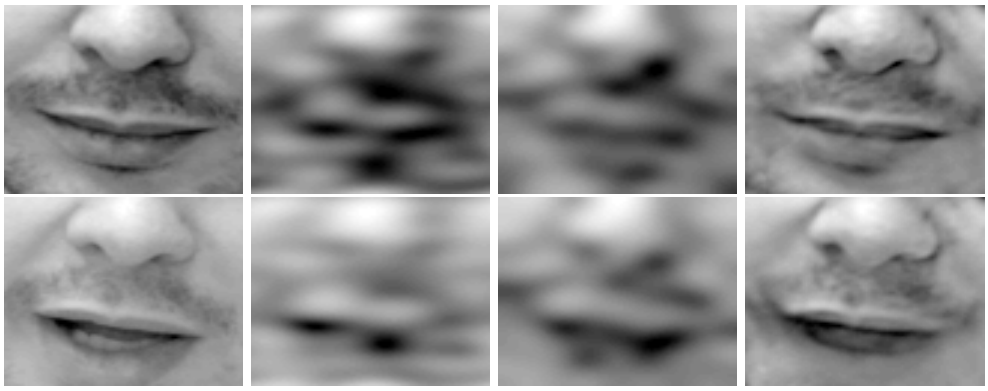

Figure 2: From left to right: Mouth region extracted from the image and converted to gray image as well as the result of the reconstruction using 50 coefficients either with FFT, DCT or PCA. Two different images are compared.

timal number of features and then a decrease. This optimal point is reached at 25 features for DCT, 50 features for FFT and 250 features for PCA. Due to the higher energy compaction of the DCT versus the FFT it is not surprising to see that the DCT reaches the same performance than FFT with less coefficients. Yet following this line of argument the PCA should need even less coefficients as it yields orthogonal coefficients with decreasing energy. When looking at Fig. 2 we see that for the closed mouth the PCA retained even with 50 coefficients a lot of detail from the original image. On the other hand when looking on the reconstruction from the open mouth we also see that the PCA introduces artifacts. In the reconstruction the mouth seems almost to be closed. Hence the PCA seems not as well be able to capture the information relevant for the task as either FFT or DCT. FFT and DCT mainly extract the low frequency components which correlate well with mouth opening and spreading.

The decrease in performance when surpassing a certain number of coefficients used can most likely be explained by the fact that with the limited amount of training data the SVM is not able to separate the features well in higher dimensional spaces. Surprisingly this is not the case for the PCA. Partly this might be explained by the orthogonality of the features, yet due to the estimation of the covariance matrix on a smaller subset this orthogonality is not perfect.

Adding first and second derivatives to the features leaves the general picture mostly unchanged but rather has a negative effect on the performance.

The combination of the visual features with acoustic features shows that the visual information is complementary to the acoustic information. When the acoustic energy is combined with either image transformation feature results statistically significantly improve for almost all feature dimensionalities (from 61% to $\approx$ 68%). In this case the performance of either acoustic energy or the image transformation features was comparable The combination of energy and fundamental frequency yields much better performance (61% vs. 76%). As a consequence the improvement from adding the visual information is smaller but still statistically significant (from 76% to $\approx$ 80%).

The results varied notably for the different speaker, as well in the acoustical as in the visual modality. Such a high degree of variability has been reported in previous experiments [13, 23]. In particular this renders the automatic determination of the prominence of a word difficult.

The subjects tilted their heads quite frequently during the interaction. The current visual feature extraction does not take care of this. For this reason we expected the FFT, where the magnitude is position invariant, to perform better than the DCT. We did not see this, maybe because this effect was counterbalanced by the higher energy compaction of the DCT. Significant improvements of the visual features can be expected once this tilt is corrected.

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. 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 my subjects for their patience and effort.

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. 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] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, p. 212, 1954.

[6] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovi-

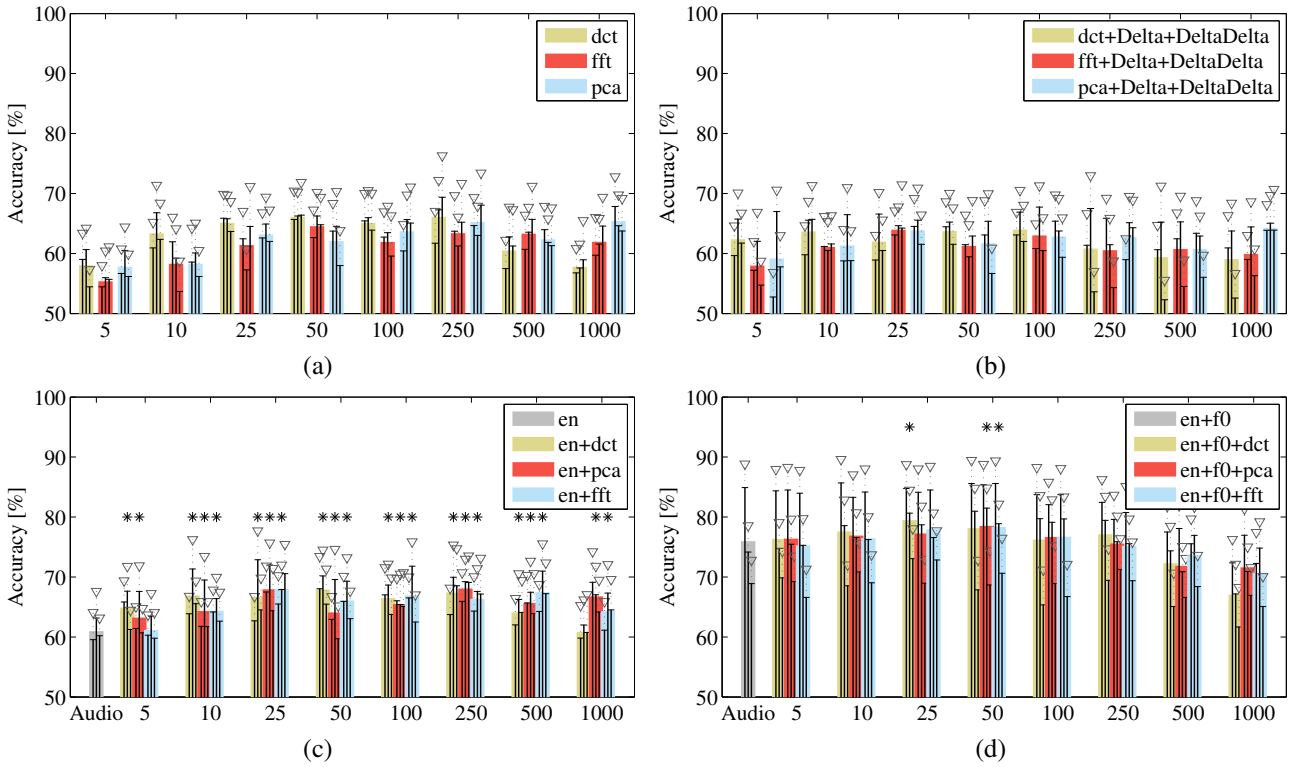

Figure 3: Discrimination accuracies using the FFT, DCT and PCA (a) including deltas and delta-deltas (b) or in combination with the relative energy (c) or relative energy and fundamental frequency (d). The number of features retained from the transformation is indicated on the abscissa. The wide and colored lines represent the mean value of all three speakers tested. The small black lines are the results of the individual speakers where the small triangles indicate the standard deviation of the 30 fold cross validation underlying each data point. Combinations of acoustic and visual features which yielded statistically significantly better results than the corrsponding acoustic features are marked with an asterisk (significance was evaluated with a two-sided t-test and $p < 0.05$).

sual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[7] 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.

[8] 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.

[9] 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*, pp. 604–609, IEEE, 2009.

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

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

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

[13] 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.

[14] E. Cvejic, J. Kim, and C. Davis, "Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion," *Speech Communication*, vol. 52, no. 6, pp. 555 – 564, 2010.

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

[16] A. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," tech. rep.

[17] 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, p. 2421, 2006.

[18] M. Heckmann, X. Domont, F. Joublin, and C. Goerick, "A hierarchical framework for spectro-temporal feature extraction," *Speech Communication*, vol. 53, no. 5, pp. 736 – 752, 2011. Perceptual and Statistical Audition.

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

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

[22] C.-C. Chang and C.-J. 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`.

[23] 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.