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

Optical Flow Field Features for Audio-visual Word Prominence Detection

Andrea Schnall, Martin Ernst Heckmann

2015

Preprint:

This is an accepted article published in Int. Joint Conf. on Neural Networks (IJCNN). The final authenticated version is available online at: https://doi.org/[DOI not available]

Evaluation of Optical Flow Field Features for the Detection of Word Prominence in a Human-Machine Interaction Scenario

Andrea Schnall*†,Martin Heckmann†
*TU Darmstadt - Control Methods and Robotics
Holzhofallee 38, 64295 Darmstadt, Germany
Email: aschnall@rmr.tu-darmstadt.de
†Honda Research Institute Europe GmbH
Carl-Legien-Str. 30, 63073 Offenbach/Main, Germany
Email: martin.heckmann@honda-ri.de

***Abstract*—In this paper we investigate optical flow field features for the automatic labeling of word prominence. Visual motion is a rich source of information. Modifying the articulatory parameters to raise the prominence of a segment of an utterance, is usually accompanied by a stronger movement of mouth and head compared to a non-prominent segment. One way to describe such motion is to use optical flow fields.**

During the recording of the audio-visual database we used for the following experiments, the subjects were asked to make corrections for a misunderstanding of a single word of the system by using prosodic cues only, which created a narrow and a broad focus.

Audio-visual recordings with a distant microphone and without visual markers were made. As acoustic features duration, loudness, fundamental frequency and spectral emphasis were calculated. From the visual channel the nose position is detected and the mouth region is extracted. From this region the optical flow is calculated and all the optical flow fields for one word are summed up. The pooled optical flow for the four directions is then used as feature vector. We demonstrate that using these features in addition to the audio features can improve the classification results for some speakers. We also compare the optical flow field features to other visual features, the nose position and image transformation based visual features. The optical flow field features incorporate not as much information as image transformation based visual features, but using both in addition to the audio features leads to the overall best results, which shows that they contain complementary information.

I. Introduction

Prosody plays an important role in human interaction and it could be useful for human-machine interaction as well. A problematic case in human machine conversation are corrections. When we correct a misunderstood word it often gets even worse because humans tend to use different prosody to indicate a correction [1]. Modifying the articulatory parameters to raise the prominent of a word is a signal for a correction. Detecting prominence could be used to detect presence and location of a correction word and so improve communication with a machine and make it more natural. But still, integrating such prosodic cues in human-machine interaction system is a challenging task. Consequently, so far it has been rarely used in spoken dialog systems [2], [3]. The reason for this might be that prosodic cues are difficult to extract due to their variation depending on the speaker and the difficulty to extract them from the speech signal.

In our previous work we investigated the audio-visual discrimination between prominent and non-prominent words [4]. Visual features are not yet often used for speech processing, especially not for prosody. But in head and mouth movements a lot of information is incorporated and it is an important part of human communication. Other authors [5], [6], [7], [8] already showed that automatic speech recognition can be improved by using this visual information, especially in difficult situations. The authors of [9], [10] showed how optical flow could be used to improve lip reading from images. Furthermore, it has been shown, that the visual channel also conveys relevant information for the perception of prosody [11], [12], [13] and in particular for word prominence [14], [15].

The mouth movement is more intense when emphasizing a word and often it is accompanied by a head movement and more detailed with a larger jaw opening and lip spreading [16], [17].

In this work we will investigate how the visual movement, described by optical flow fields can affect the detection of prominence. We will compare it to other visual features we used before like discrete cosine transformation based features and the nose position, and test the combination of them in addition to the audio features. All visual feature types are extracted from images of a speaker's face interacting with a computer without additional markers positioned in his face.

The next section presents the audio-visual database used, followed by a description of the audio features and the image transformation based features including the optical flow field features in section III and the used feature combinations. A short description on the classification is given in section IV. In section V the results will be presented, followed by a discussion in section VI. In the final section we give a conclusion.

II. Dataset

The audio-visual data was recorded as a Wizard-of-Oz experiment in a small game where the subjects moved tiles to uncover a cartoon [4]. Therefore, the subjects had to give spoken advises, in a simple grammar, to a computer of the form "place green in B one". Occasionally, a misunderstand-

978-1-4799-1959-8/15/$31.00 ©2015 IEEE

ing of one word of the sentence was triggered, which was verbally and visually shown. Then the subjects had to correct the sentence by repeating them and using prosodic cues to emphasize the misunderstood word as they would do with a human. However correction cue phrases like "no I said" were not allowed. This procedure should lead to a rather natural use of prosody, creating a narrow focus condition (in contrast to the broad focus condition of the original utterance) with the corrected words marked as highly prominent.
In total 16 subjects, male and female, speaking either British or American English as native language or being either bilingual British English/German or American English/German were recorded. This resulted in over 2500 utterances with 4 to 5 words for evaluation. The speech was recorded with a distant microphone at 48 kHz and later downsampled to 16 kHz.
For the video images a CCD camera with a resolution of 1280x1024 pixels and a frame rate of 25Hz was used. No visual markers were made. A speech recognition system trained on the Grid Corpus [18] was used to perform a forced alignment, which provides a temporal annotation of the data and is used in the following to determine the boundaries of the different words.
Three human annotators annotated the recorded data with a 4 level scale (0-3) of prominence for each word. We calculated the inter-annotator agreement with Fleiss' kappa κ. While doing so we binarized the annotations, i. e. only differentiating between prominent and non-prominent. A word was annotated as prominent if the mean annotation of all annotators was above 1.5. During the annotation we saw that on the one hand, not all speakers consistently used prominence to highlight the corrected word and, on the other hand, that the annotators were for some speakers not able to come to a consensus on the prominence of the words. Therefore, from the originally 16 speakers we retained 8 speakers which showed an agreement measured with Fleiss' κ of more than 0.55 between all annotators ($0.4 < \kappa < 0.6$ is usually considered as moderate agreement) and which are used in the following experiments.

III. Features

In this section we will describe the used features, audio and visual, respectively the feature-combinations.

A. Audio Features

From the audio-channel the features which have been previously proposed to correlate with word prominence are used. From these features the mean value for each word is calculated and used in the subsequent analysis (except for duration and gap length). The beginning and end of the word is taken from the forced alignment.
We then calculate as audio features the duration of a word (D_W) as well as the gap length before and after the word as determined from the forced alignment (D_G).
The loudness (l) is calculated by first filtering the signal with an 12th order IIR filter following the ideas outlined in [19], next calculating the instantaneous energy, followed by smoothing with a low pass filter with a cut-off frequency of 10 Hz, and finally converting the results to dB. We expected the loudness to better capture the perceptual correlated of prominence than the energy.

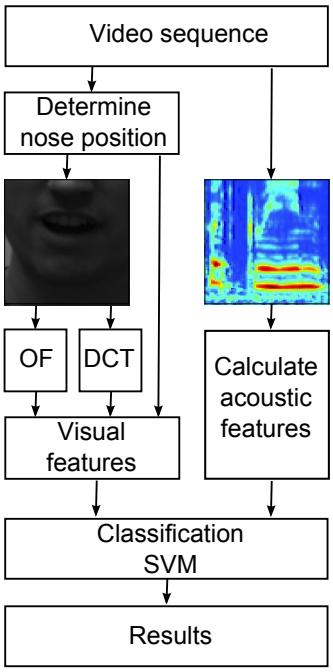


Fig. 1. From the video sequence, on the one hand a range of different audio features is calculated, on the other hand the nose position of the speaker is determined and with this knowledge the DCT transformation and the optical flow of the mouth region. All features are than used for the classification with support vector machines.

Another feature we use is the 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.
As described in [20], we extract the fundamental frequency f_0 (following [21]), interpolate values in the unvoiced regions via cubic splines and convert the results to semitones. To detect voicing we use an extension of the algorithm described in [22].

From this acoustic features, we extract functionals as described in [21] for a better prosodic analysis. We extract the mean, max, min, spread (max-min) and variance along the word. Hereby word boundaries are determined by the forced alignment. Prior to the calculation of the functionals we normalize the prosodic features by their utterance mean and calculated their first and second derivative (except for duration and gap length).

B. Visual Features

As visual features we use the nose position and the discrete cosine transformation features, and the features based on optical flow.

1) Nose Position and DCT Features: From the visual channel, the region around the mouth is detected. Therefore we start by detecting the nose in each frame with the help of the openCV library [23]. We detect the position of the nose (Nose), instead of detecting the mouth directly, since the shape of the nose does not change. Also the nose itself does not move much relative to the skull while talking so it is a good measure for the rigid head movements.
Based on the nose position we detect the eye's position and use it for calculating the head tilt angle and compensating for it by

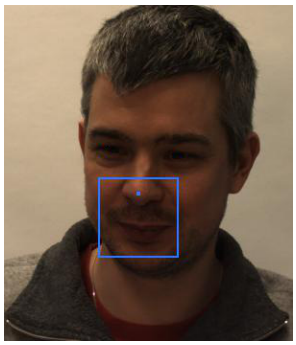


Fig. 2. Image from recording after cropping to face region, nose detection, down sampling and highlighting of the mouth region.

rotating the image. With the knowledge of the position of the nose, the mouth position can be calculated by using a fixed speaker independent offset as depicted in Fig.2. The mouth region is centered by calculating the symmetry axis using the algorithm proposed in [24]. The cropped mouth region has, in our case, after down sampling by a factor of 2, a 100x100 pixel size. From this region a two-dimensional discrete cosine transform (DCT) is calculated. Out of the 10000 coefficients per image we select the 50 with the highest energy [25].

2) Optical Flow Field Features: To describe movements in Terms of direction and intensity the optical flow is a possible choice. Optical flow describes the movement of brightness patterns in an image. For calculating the optical flow we use the classical formulation by Horn and Schunck (OFHS) [26] as well as the state-of-the-art segmentation-based optical flow field method by Sun et al. (OFSun) [27]. The classical formulation of the objective function, as written in [27], is in its spatially discrete form:

$$
\begin{aligned}
&E(\mathbf{u},(u,v)) = \\
&\quad \sum_{i,j} \{\rho_D(I_1(i,j) - I_2(i+u_{i,j}, j+v_{i,j}) \\
&+\lambda\left[\rho_s(u_{i,j}-u_{i+1,j}) + \rho_s(u_{i,j}-u_{i,j+1})\right. \\
&\quad \left. + \rho_s(v_{i,j}-v_{i+1,j}) + \rho_s(v_{i,j}-v_{i,j+j})\right]\},
\end{aligned}
$$

where u and v are the horizontal and vertical components, I_1 and I_2 are the succeeding images, λ the regularization parameter, ρ_D the data penalty function and ρ_S spatial penalty function. In our experiments we used the Charbonnier penalty function (method 'classic-c'):

$$
\rho(x) = \sqrt{x^2+\epsilon^2},
$$

with ϵ being a small number, e.g. 0.001.
For the optical flow field method by Sun et al. a non-local term is added to the flow energy of the classical formulation:

$$
\begin{aligned}
&E(\mathbf{u},\mathbf{v},\hat{\mathbf{u}},\hat{\mathbf{v}}) = \\
&\quad \sum_{i,j} \{\rho_D(I_1(i,j) - I_2(i+u_{i,j,j+v_{i,j}}) \\
&+\lambda\left[\rho_s(u_{i,j}-u_{i+1,j}) + \rho_s(u_{i,j}-u_{i,j+1})\right. \\
&\quad \left. + \rho_s(v_{i,j}-v_{i+1,j}) + \rho_s(v_{i,j}-v_{i,j+j})\right]\} \\
&+\lambda_2(\|\mathbf{u}+\hat{\mathbf{u}}\|^2 + (\|\mathbf{v}+\hat{\mathbf{v}}\|^2) \\
&+\sum_{i,j} \sum_{(i',j')\in N_{i,j}} w_{i,j,i',j'}(\|\hat{u}_{i,j}-\hat{u}_{i',j'}\| + \|\hat{v}_{i,j}-\hat{v}_{i',j'}\|),
\end{aligned}
$$



(a) (b) (c)

Fig. 3. (a) Mouth region, (b) optical flow calculated with the classical formulation by Horn & Schunk [26], and (c) calculated with the formulation of Sun et al. [27].

where $\hat{u}$ and $\hat{v}$ stand for the vertical and horizontal component of an auxiliary flow field and $N_{i,j}$ is the set of neighbors of pixel (i,j) and λ_2 is a scalar weight. $w_{i,j,i',j'}$ weights the pixels depending on their probability to belong to the same surface. (For both methods we use the publicly available implementation described in [27]). For the calculation the default parameters are used.
Compared with the Horn and Schunck method the method of Sun et al. is capable of preserving more amount of detail due to the local segmentation as can bee seen in Fig. 3.
For the classification the optical flow field for the mouth region in all frames is calculated. Afterwards the optical flow fields belonging to one word are summed up. Then the image is divided into different regions, e.g. 4 times 5 regions, and for every region the values for every direction is summed up. The resulting feature vector is then used for classification.

C. Feature Combinations

For classification we test different combinations of the features. Therefore the different features are stacked to one features vector before classifying. We investigate the combination of all acoustic features, including loudness, duration, gap length, spectral emphasis and the fundamental frequency. Also we test the visual features alone, i.e. the optical flow field calculated using the classical formulation of Horn and Schunck and the version of Sun et al., the DCT features calculated from the mouth region and the 2D nose position.
And finally we test the different combinations of those features, i.e. the acoustic features with the different visual features and the combination of all features:

1) The single visual features:
   OFHS, OFSun, Nose, DCT
2) The combination of all visual features:
   OFHS/OFSun-Nose-DCT

3) The combination of all audio features (Audio): l-D_W-D_G-Se-f_0
4) The different combinations of audio and video features:

   a) Audio-OFHS
   b) Audio-OFSun
   c) Audio-Nose-DCT
   d) Audio-OFHS-Nose-DCT
   e) Audio-OFSun-Nose-DCT.

D. Context Features

As described in [28], based on the hypothesis that emphasizing a word by hyperarticulating leads to a hypoarticulation of the neighboring words, we analyzed the performance of the single words as well as the performance of features based on the knowledge of the neighboring words. That means that in addition to the features of the word itself we use the previous and the following word for classification, in the following refered as context features.

IV. Support Vector Machines

For classification we chose the support vector machines (SVM) since they show good performance while having low computational cost in comparison with other methods like Gaussian mixture models [29] and conditional random fields [28]. For classification with support vector machines (SVM) we use the libsvm [30] with a radial basis function kernel. For the classification, the data was normalized to the range [-1,1]. For the parameter C, the penalty parameter of the error term, and the variance scaling factor of the basis function, a grid search on the whole dataset was performed for every feature combination.
Then we divided our data in a training set, containing 75% of the data, and a test set, containing the remaining 25% of the data. 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. The training and testing was done for every speaker independently. As features we used the different combinations described in section III-C.

We then extended the SVM features with the information of the neighboring segments as described in section III-D. The SVM are then again trained as described above with the context-features.

Compared to the detection of pitch accents, prominent words occur less often. While e.g. in [31] the classes are nearly balanced (approx. 60% unaccented), in our database, there are approx. 8 times more non-prominent words than prominent words. Because of this unbalanced database, a comparison of the accuracy is not suitable; a preference of non-prominent words would always lead to very good accuracy but would not provide an objective evaluation. Therefore, we use the unweighted accuracy respectively the unweighted error rate

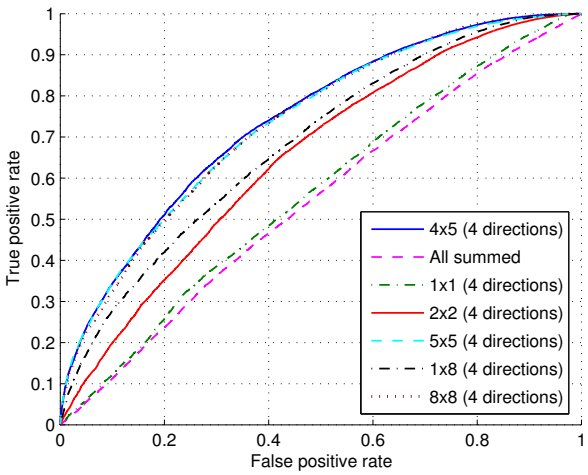


Fig. 4. Receiver operating curve of different pooling of the optical flow, averaged over 8 speaker.

TABLE I. Equal error rate of the different pooling regions for OFHS

	average over 8 speaker	best speaker
1x1	47.9	43.3
2x2	42.8	35.9
4x5	34.7	23.7
5x5	37.0	24.7
1x8	41.2	27.8
8x8	36.3	22.4

for evaluating the performance of the prominent words:

$$\text{recall} = \frac{tp}{tp+fn}, \quad (1)$$

$$\text{specificity} = \frac{tn}{tn+fp}, \quad (2)$$

$$\text{unweighted error rate} = 1 - \frac{\text{recall} + \text{specificity}}{2}, \quad (3)$$

with tp = true positive, tn = true negative, fp = false positive and fn = false negative.

V. Results

This section presents all results of the unweighted error of non-prominent and prominent words, using all words of the utterances (4 to 5 words). For the SVM different operating points result from varying the decision threshold of the classification confidence. Based on the probability estimates, calculated by the SVM, a receiver operating curve, with the true positive rate on the y-axis and false positive rate on the x-axis, is calculated as described in [32].

Fig. 4 and Table I display the results of the classification, using only the optical flow field features, for different pooling regions, varying from 1 to 8 regions in vertical and horizontal direction. Using only the sum of the whole movement between two pictures, implying that only the amount of movement is sufficient to distinguish between a prominent and non-prominent word, leads to poor results. Considering every direction of movement for itself improves the classification,

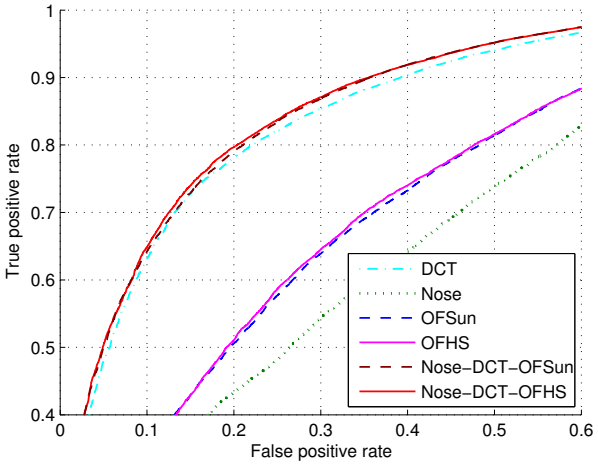


Fig. 5. Receiver operating curve of the different visual features, and the combination of all of them.

TABLE II. EQUAL ERROR RATE OF THE DIFFERENT VISUAL FEATURES AVERAGED OVER 8 SPEAKERS IN %

Nose	38.6
OFHS	34.7
DCT	21.1
Nose-DCT-OFHS	21.0

TABLE III. EQUAL ERROR RATE FOR ALL COMBINATIONS, SINGLE FEATURES

Average over 8 speakers		+ OFSun	+ OFHS
Audio	15.0	14.7	14.5
Audio-Nose-DCT	13.6	13.2	13.2
One speaker		+ OFSun	+ OFHS
Audio	14.73	13.05	11.86
Audio-Nose-DCT	11.43	10.95	11.86

but it is still only slightly over chance level. Dividing the image in different regions and summing the movement for every direction results in an unweighted error of 34.7% . The best performance was found for 4 regions in x-direction and 5 in y-directions. Using only the horizontal or vertical direction or dividing the image only in x or y direction as proposed in [9] worked not as good for our problem. The results are only shown for the optical flow using OFHS, since the results were quite similar for OFSun.

Fig. 5 and Table II shows the classification result of using only the visual features, comparing the optical flow (OFHS and OFSun), the DCT of the mouth region and the 2D nose position, as well as all of them. It can be seen that the optical flow field features worked better than the nose position but not as good as the DCT features. A bit better was the combination of all of them with a result of 21.0 %.

Fig. 6 shows the results of the feature combination (III-C) while Fig. 8 shows the results of the same combinations but using them as context features. The corresponding unweighted error can be found in Table III respectively Table IV.
Both figures shows the results for combining the Audio features and the Audio+DCT+Nose with optical flow.

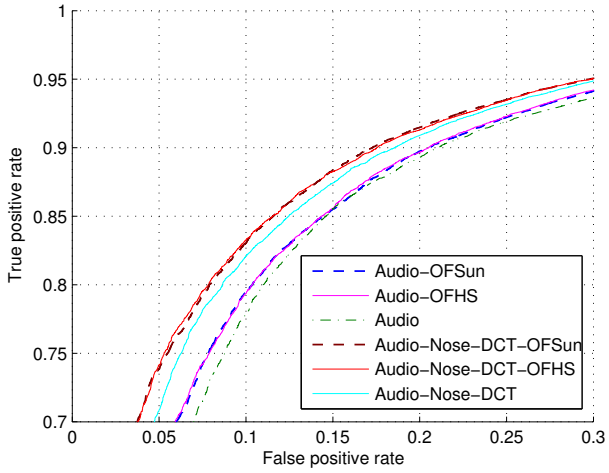


Fig. 6. Receiver operating curve of different combinations of audio and video features using single features, averaged over 8 speaker

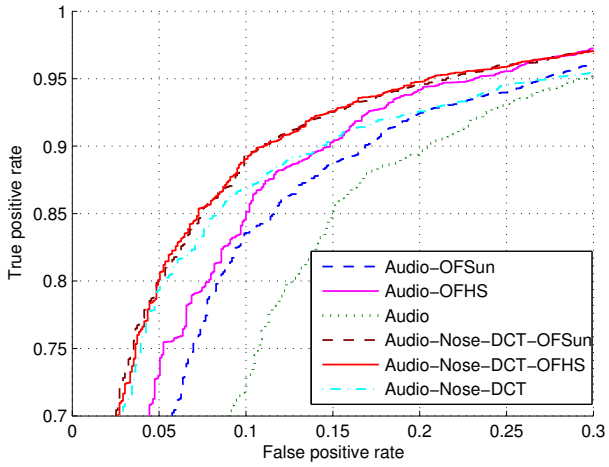


Fig. 7. Receiver operating curve of different combinations of audio and video features using single features, for 1 speaker.

TABLE IV. EQUAL ERROR RATE OF ALL COMBINATIONS, USING CONTEXT FEATURES IN %

Average over 8 speakers		+ OFHS
Audio	11.2	11.9
Audio-Nose-DCT	12.4	12.6
One speaker		+ OFHS
Audio	12.5	11.0
Audio-Nose-DCT	12.6	12.8

Comparing the results of OFHS and OFSUN the results for OFHS are a bit better. It can be seen, for single features, that adding the optical flow field features improves the results. Overall the improvement through adding visual features to the audio features is 1.8%

VI. DISCUSSION

As discussed in former publications, using visual features can reduce the detection error.

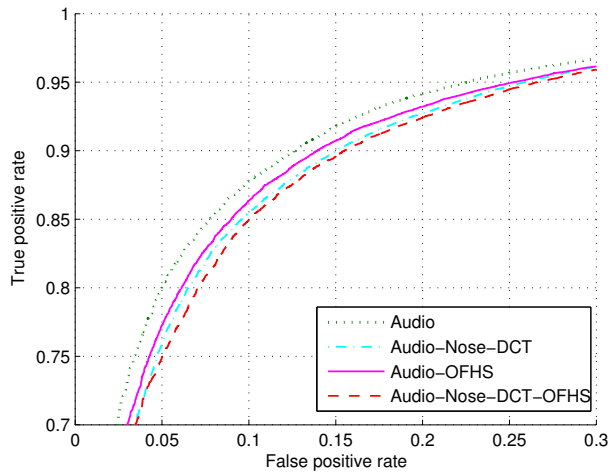


Fig. 8. Receiver operating curve of different combinations of audio and video features using context features, averaged over 8 speaker.

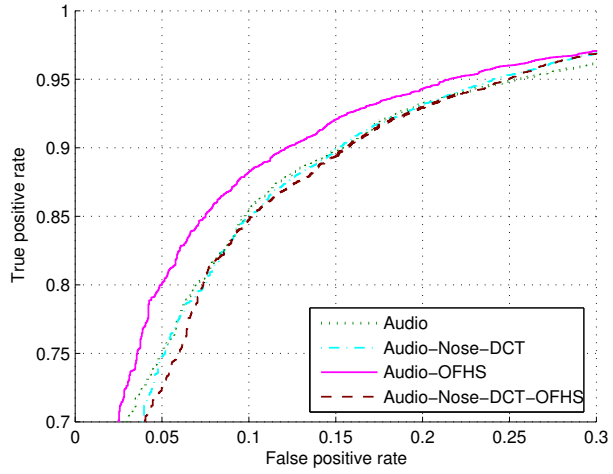


Fig. 9. Receiver operating curve of different combinations of audio and video features using context features, for 1 speaker.

Our results demonstrate that using optical flow field features additional to the audio features in general improves the results. Using the classical formulation of optical flow from Horn & Schunck works in our case better than the formulation of Sun et al., which shows that a detailed preservation of the optical flow is not helpful for our problem. This might be because the difference of prominent and non-prominent words lies more in the intensity of the movement than in the actual shape. But the performance difference between both is not large.
But the experiments on dividing the flow in the different directions and regarding different regions of the mouth showed, that not only the intensity is effecting the classification results. The optical flow field features in comparison to the nose position and DCT features contain less information, but using all of them in addition to the audio features leads to the best performance, which shows that the optical flow field features contain complementary information to the other features.
Also the results for different speakers were quite different. For some speakers the visual features did significantly improve the performance while for others they were not helpful. For one speaker the improvement was from 14.7% to 11.8%. Regarding the results of context features, there the overall performance is better. There the acoustic features already contain more information than the single features and the visual features can not further improve the performance. Also, for the context features, the Audio+OFHS performance is better than Audio+DCT+Nose.
The impact of visual features is very different for the different speakers, Fig. 7 and 9 shows the roc for one speaker and single features respectively context features, where the optical flow fields capture notable more information. For this speaker the results improve from 14.7% to 11.8% and 12.5 to 11%.
In our Database for 5 speakers the classification could be improved by adding optical flow to audio features while for 3 speakers the error increases. Adding optical flow to the combination of Audio+Nose+DCT still improves the results for 4 speakers.

The improvement through visual features might not seem that big, but could gain importance in case of noisy environments where the audio features are not as stable as the visual features. The best result we obtained for single features was using all audio and visual features with 13.2% unweighted error. For context features the best obtained results where 12.5% for using only the audio features.

VII. Conclusion

We tested optical flow field as a feature for the detection of prominent words in speech on a database with 8 subjects. Our results showed that the optical flow field features in addition to the audio features reduced the error for the detection of prominent words. Comparing it with our former visual features, they contain less but complementary information that are not yet included. The visual features might become even more important in case of noisy surrounding where the audio features are not that reliable anymore.
Overall the best performance was reached by using the audio features as well as nose position, the DCT features and the optical flow fields features.

Acknowledgment

We thank Heiko Wersing and Petra Wagner for fruitful discussions. Many thanks to Mathias Franzius for support with tuning the SVMs and Merikan Koyun for help in the data preparation.

References

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

[2] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van, and E. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, pp. 339–373, 2000.

[3] J. Hirschberg, D. Litman, and M. Swerts, "Characterizing and predicting corrections in spoken dialogue systems," *Comput. Linguist*, vol. 32, pp. 417–438, 2006.

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

[5] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. 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. Orglmeisterm, "Audiovisual speech recognition with missing or unreliable data," *Proc.Int. Conf.Auditory Visual Speech Process. (AVSP)*, 2009.

[8] 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. IEEE*, p. 604609, 2009.

[9] A. Shaikh, D. Kumar, W. Yau, M. Che Azemin, and J. Gubbi, "Lip reading using optical flow and support vector machines," in *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol. 1, Oct 2010, pp. 327–330.

[10] S. Tamura, K. Iwano, and S. Furui, "Multi-modal speech recognition using optical-flow analysis for lip images," in *Real World Speech Processing*, J. Wang, S. Furui, and B. Juang, Eds. Springer US, 2004, pp. 43–50.

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

[12] 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.

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

[14] 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.

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

[16] E. Cvejic, J. Kim, C. Davis, and G. Gibert, "Prosody for the eyes: Quantifying visual prosody using guided principal component analysis," in *INTERSPEECH. ISCA*, 2010.

[17] M. Dohen, H. Loevenbruck, and H. Harold, "Visual correlates of prosodic contrastive focus in french: Description and inter-speaker variability," in *Speech Prosody, Dresden, Germany*, 2006.

[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[19] "Replaygain 1.0 specification," http://wiki.hydrogenaudio.org/.

[20] M. Heckmann, "Steps towards more natural human-machine interaction via audio-visual word prominence detection," in *2nd Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction, Singapore*, 2014.

[21] M. Heckmann, F. Joublin, and C. Goerick, "Combining rate and place information for robust pitch extraction," in *INTERSPEECH, Antwerp*, 2007, p. 27652768.

[22] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," *INTERSPEECH*, vol. 2, p. 3, 2005.

[23] G. Bradski, "The opencv library," *Dr. Dobb's Journal of Software Tools*, 2000.

[24] M. Nishigaki, S. Rebhan, and N. Einecke, "Vision-based lateral position improvement of RADAR detections," in *Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2012, pp. 90–97.

[25] M. Heckmann and A. Schnall, "Effects of background noise on audio-visualword prominence detection," submitted to *INTERSPEECH, Dresden*, 2015.

[26] B. Horn and B. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 16, pp. 185–203, 1981.

[27] D. Sun, S. Roth, and M. Black, "Secrets of optical flow estimation and their principles," *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2439, 2010.

[28] A. Schnall and M. Heckmann, "Integrating sequence information in the audio-visual detection of word prominence in a human-machine interaction scenario," in *INTERSPEECH, Singapour*, 2014.

[29] ——, "Speaker adaptation for word prominence detection," submitted to *INTERSPEECH, Dresden*, 2015.

[30] 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.

[31] M. Gregory and Y. Altun, "Using conditional random fields to predict pitch accents in conversational speech," in *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, 2004, pp. 677–683.

[32] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," Tech. Rep., 2004.