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

Feature-level Decision Fusion for Audio-Visual Word Prominence Detection

Martin Ernst Heckmann

2016

Preprint:

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

INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Feature-level Decision Fusion for Audio-visual Word Prominence Detection

Martin Heckmann

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

Abstract

Common fusion techniques in audio-visual speech processing operate on the modality level. I.e. they either combine the features extracted from the two modalities directly or derive a decision for each modality separately and then combine the modalities on the decision level. We investigate the audio-visual processing of linguistic prosody, more precisely the extraction of word prominence. In this context the different features for each modality can be assumed to be only partially dependent. Hence we propose to train a classifier for each of these features, acoustic and visual modality, and then combine them on a decision level. We compare this approach with conventional fusion methods, i.e. feature fusion and decision fusion on the modality level. Our results show that the feature-level decision fusion clearly outperforms the other approaches, in particular when we also additionally integrate the features resulting from the feature fusion. Compared to a detection based only on the full audio stream we obtain relative improvements from the audio-visual detection of 19% for clean audio and up to 50% for noisy audio.
Index Terms: audio-visual, fusion, naïve Bayesian, prominence, background noise, prosody

1. Introduction

A large body of research has investigated the benefits of combining acoustic and visual information for audio-visual speech recognition [1, 2, 3, 4, 5]. However, the research on the audio-visual processing of prosody only focuses on emotional prosody [6]. Previously it had been shown that humans are able to use visual information to extract prosodic cues [7, 8, 9, 10]. Yet there is to our knowledge no research into the benefits of visual information in the context of the classification of linguistic prosodic events. The system we presented in [11] was to our knowledge the first such system. In this system we investigated the audio-visual detection of prominent words. Humans use prosodic cues to highlight a correction after a misunderstanding when talking to another human but also when talking to a machine [12]. A distinguishing feature of corrections is that they are frequently hyperarticulated and hence very prominent [13]. Acoustic correlates of prominence have been shown to include a longer duration as well as specific pitch and intensity patterns [14, 15]. In the visual modality prominence is mainly manifested in larger jaw opening, lip spreading and protrusion and to some extent via head movements [16, 17]. Prominence and pitch accent are terms which frequently occur together. One approximation to their relation is that perceived prominence results from a pitch accent [18]. Different methods have been developed to detect pitch accent and prominent words from the acoustic modality [19, 13, 20].

In [21] we extended our previous system and evaluated it also when noise was present in the acoustic modality. In this paper we have a closer look at the fusion of the audio and video modality. We investigate previously presented fusion methods which are mainly based on a feature-level or decision-level fusion and introduce the feature-level decision fusion.

In the next section we describe the different fusion methodologies. Following this we introduce the dataset we used for our experiments. We detail the different features extracted from the acoustic and visual channel in Section 3. Section 4 will present the results of our experiments. After that we will discuss the results in Section 5 and give a conclusion in Section 6.

2. Audio visual fusion

In the context of audio-visual speech recognition and multimodal fusion in general many different approaches for the fusion of the audio and video stream have been proposed [1, 22]. These fusion methods are mainly classified as feature-level and decision-level fusion. In the next section we will give more details on these approaches and also introduce our novel feature-level decision fusion which is based on a naïve Bayesian model.

2.1. Feature Fusion

We implemented the feature fusion (FF), also called feature concatenation, as [1]:

$$\mathbf{o}_{AV} = [\mathbf{o}_A, \mathbf{o}_V] \in \mathbb{R}^{l_{AV}}, \text{ where } l_{AV} = l_A + l_V.$$

Hence we concatenated the feature vectors of the two modalities to a larger feature vector.

2.2. Decision Fusion

For decision fusion (DF) we performed a classifier combination. I. e. we individually classified the two modalities and then fused the decisions. As decisions we use a posteriori probabilites $P(C_i|\mathbf{o})$ provided an SVM which we use for classification. While doing so we assume class conditional independence between the two modalities:

$$P(\mathbf{o}_A, \mathbf{o}_V|C_i) = P(\mathbf{o}_A|C_i) * P(\mathbf{o}_V|C_i),$$

where C_i represents the class i, in our case prominent or non-prominent. Using Bayes formula on derives at [2]:

$$P(C_i|\mathbf{o}_A, \mathbf{o}_V) = \frac{P(C_i|\mathbf{o}_A) * P(C_i|\mathbf{o}_V)}{P(C_i)} * \eta(\mathbf{o}_A, \mathbf{o}_V)$$

where $P(C_i)$ represents the a priori probability of class i. The normalization term $\eta(\mathbf{o}_A, \mathbf{o}_V)$ is independent of the class C_i and can hence be neglected for the classification.

2.3. Feature-level Decision Fusion

For the decision-level fusion conditional independence at the modality level is assumed. In the machine learning community

Copyright © 2016 ISCA

http://dx.doi.org/10.21437/Interspeech.2016-163

a common approach is called a naïve Bayesian model which assumes independence on the level of the different features used for the classification [23]. Given N features this yields:

$$P(C_i|\mathbf{o}_1,\ldots,\mathbf{o}_N) = \frac{\prod_{n=1}^{N} P(C_i|\mathbf{o}_n)}{P(C_i)^{N-1}} * \eta(\mathbf{o}_1,\ldots,\mathbf{o}_N)$$

Such naïve Bayes classifiers generally perform well even if the underlying assumption is not met. One reason for this is that the assumption of conditional independence is a sufficient but not a necessary condition. In many cases, in particular when the dependencies of the different features are equally distributed amongst the classes or cancel each other out, the naïve Bayes classifier has equal performance to the optimal classifier [23].

In the context of noise robust speech recognition an equivalent approach to the naïve Bayesian model was developed and termed Full Combination Approximation [24]. This approach was motivated by observations that in human perception, the error rate for fullband perception is approximately equal to the product of the sub-band error rates obtained when the each subband was perceived on its own [24]. To obtain good performance in the Full Combination Approximation also the fullband posteriors were included. We follow this approach by including the features derived from feature fusion of the individual modalities and both modalities at the same time. We termed this as full feature-level decision fusion. Additionally we included features derived from fusing only a subset of the features of one modality on the feature level. We termed this as extended feature-level decision fusion.

2.4. Weighted Fusion

In particular in audio-visual speech recognition it has been observed that weighting the different modalities during fusion depending on their reliability largely improves performance [1, 2, 3]. We also investigated the following weighting in our experiments:

$$P(C_i|\mathbf{o}_{A_1},\ldots,\mathbf{o}_{A_L},\mathbf{o}_{V_1},\ldots,\mathbf{o}_{V_K},\mathbf{o}_{AV}) = \frac{\left[\prod_{l=1}^{L} P(C_i|\mathbf{o}_{A_l})\right]^{\lambda}\left[\prod_{k=1}^{K} P(C_i|\mathbf{o}_{V_k})\right]^{1-\lambda} P(C_i|\mathbf{o}_{AV})}{P(C_i)^{L+K-1}P(C_i)} * \eta(\mathbf{o}_{A_1},\ldots,\mathbf{o}_{A_L},\mathbf{o}_{V_1},\ldots,\mathbf{o}_{V_K},\mathbf{o}_{AV})$$

This means we first fused the different features of the two modalities individually according to the feature-level decision fusion and then weighted each modality. Thereby L and K are the number of features we used in the audio and video modality respectively. We also used a conventional weighted decision fusion. In this case we performed the decision fusion as outlined above, i. e. we weighted the features resulting from the feature fusion. Thereby we set $K = L = 1$ and removed the terms depending on $\mathbf{o}_{AV}$ and the final multiplication with $P(C_i)$.

3. Dataset

To stimulate corrections and hence prominent words, we recorded subjects interacting via speech in a Wizard of Oz experiment with a computer in a small game where they moved tiles to uncover a cartoon [11]. This game yielded utterances of the form 'place green in B one'. Occasionally, a misunderstanding of one word of the sequence was triggered and the corresponding word highlighted, verbally and visually. 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 16 native English speaking subjects were recorded [25]. 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.

We trained HTK [26] on the Grid Corpus [27] followed by a speaker adaptation with a Maximum Likelihood Linear Regression (MLLR) step with a subsequent Maximum A-Posteriori (MAP) step to perform a forced alignment of the data.

Three human annotators annotated the recorded data with 4 levels 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. We tested different binarizations and used the one where the agreement between all annotators was highest. Next we calculated κ for each speaker individually. We then discarded all speakers where κ for the optimal binary annotation was below 0.5 ($0.4 < \kappa \leq 0.6$ is usually considered as moderate agreement). We have chosen such a rather low threshold to retain as many speakers as possible. This yields 11 speakers, 6 females and 5 males. Overall we have 4622 utterances of which 1892 are corrections, i. e. on average ≈ 420 utterances per speaker with $\approx 40\%$ corrections.

4. Features

Most approaches in the computational processing of prosody rely on functionals derived from low level acoustic descriptors [28]. In the following we will detail which acoustic, or in our case also visual, low level descriptors, we used and which functionals we derived from them. These functionals then serve as features for a Support Vector Machine (SVM) based classifier.

4.1. Acoustic low level descriptors

Since we expected the loudness l to better capture the perceptual correlates of prominence than the energy, we extracted it by filtering the signal with an 11th order IIR filter as described in [29], followed by the calculation of the instantaneous energy, smoothing with a low pass filter with a cut-off frequency of 10 Hz, and conversion into dB. Furthermore, we calculated D, the duration of the word and the gaps before and after the word as determined from the forced alignment. We also extracted the fundamental frequency f_0 (following [30]), interpolated values in the unvoiced regions via cubic splines and converted the results to semitones. To detect voicing, we used an extension of the algorithm described in [31]. Finally, we also determined the spectral emphasis SE, i. e. the difference between the overall signal energy and the energy in a dynamically low-pass-filtered signal with a cut-off frequency of $1.5 f_0$ [32].

4.2. Visual low level descriptors

To extract features from the visual channel, we used the OpenCV library [33] to detect the face and the nose in the image. The nose does not move much during articulation relative to the head and is hence well suited to measure the rigid head movements. As the detection of the nose with OpenCV was not very reliable we implemented several post-processing steps. First we extracted two nose hypotheses for each frame and kept

those which were more plausible with respect to their position in the face. In the sequence of nose positions we looked for a temporal context where the nose position did not change much. At the center of this temporal zone we cut out a region in the image around the nose and used it as template for a correlation based nose tracking. I. e. forward and backward in time from this region we tracked the nose by correlating the current image with the nose template and determining the shift. Once we obtained the nose tracks we also determined the eyes in the image. For doing so we detect the darkest spot in the image where we expect the eyes based on the nose position, a frequently used technique [34]. Based on the eyes' position we calculated the head tilt angle and compensated for it by rotating the image. We cropped an image around the expected mouth region in the rotated image (again based on the nose position) and centered the mouth region in it by calculating the symmetry axis using the algorithm proposed in [35]. Next we cropped the actual mouth region and calculated a two-dimensional Discrete Cosine Transform (DCT) on each subsampled mouth image of size 100×100 pixels. Out of the 10000 coefficients per image we selected the 20 with the lowest spatial frequencies.

4.3. Functionals and Contours

Prior to the calculation of the functionals, we normalized the prosodic features by their utterance mean and calculated their first and second derivative (except for D). As functionals we extracted the mean, max, min, spread (max-min) and variance along the word. Word boundaries were determined by the forced alignment.

To extract additional information we also model the contours of the features via a DCT. This is a frequently used and computationally very simple method typically yielding good results [36]. Effectively the DCT transforms the contour into a frequency representation. By retaining only the K lowest DCT coefficients we represent only the low frequency variations. For the acoustic modality we set $K_A = 10$. Due to the much lower sampling rate of the visual features (25 Hz as compared to 100 Hz) we retained only $K_V = 7$ coefficients for the visual modality.

4.4. Context Features

Marking the focus of a word in an utterance, rendering it prominent, also has an influence on the neighboring words, i. e. the word in focus is hyperarticulated and the surrounding words are hypoarticulated [37, 38]. It has been shown previously that taking this context information into account is very effective for the detection of word prominence [39, 40, 19]. Therefore, we also apply this in our approach by stacking features prior to classification such that they contain not only the functionals of the current but also of the previous and following word (see [39] for details).

5. Results

To discriminate prominent from non-prominent words, a Support Vector Machine (SVM) with a Radial Basis Function Kernel was trained using LibSVM [41]. 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 $[-1 \ldots 1]$. 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 was run. To establish the 30 sets, a sampling with replacement strategy was applied where the number of elements from the prominent and non-prominent class was set corresponding to their respective frequency in the dataset. This process was performed individually for each speaker, hence all results are speaker-dependent.

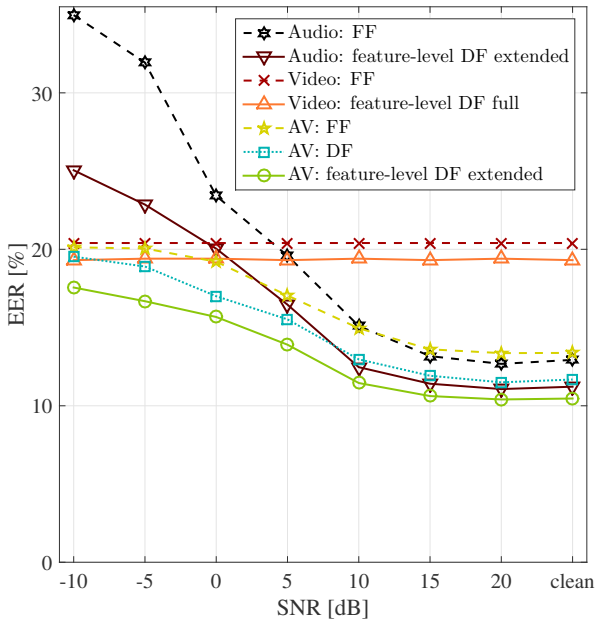


Figure 1: Equal Error Rates (EER) averaged over all 11 speakers with varying SNR levels and varying fusion approaches (FF=Feature Fusion, DF=Decision Fusion).

As proposed in [42], we calculated Receiver Operating Curves (ROC) by pooling the results of all cross-validations and all speakers. The ROC is well suited to our very unbalanced dataset where non-prominent words are approximately ten times more frequent than prominent words. From the ROC we calculated the Area Under the Curve (AUC) and the Equal Error Rate (EER). In our results AUC and EER strongly correlate so we will in the following mainly report EER.

To investigate the behavior of the fusion of the two modalities when the acoustic channel has varying reliablity we added to the clean audio signal "car" noise taken from the Noisex database [43] with varying Signal to Noise Ratio (SNR) levels using the tool Fant [44]. We trained and tested the SVMs for identical SNR, i. e. we performed a matched training. Thereby we always kept the alignment obtained from the clean signals. This will overall lead to results which are most certainly better than in a real environment. Yet it avoids the dependence of the results from the performance of the speech recognition system in noise. With respect to our audio-visual fusion experiments it will rather lead to a pessimistic estimation. In general larger gains are expected from the audio-visual fusion when the audio modality is less reliable.

For the extended feature-level decision fusion we also included two features from the audio stream which were based on the feature fusion of all acoustic features without either f_0 or loudness.

To calculate the weighted fusion we varied the parameter λ and determined for each SNR level the λ value yielding the best results averaged over all speakers.

Features	clean		with noise	
	AUC	EER in %	AUC	EER in %
Audio				
FF	0.943±0.005	12.9%	0.859±0.008	20.5%
feature-level DF	0.948±0.005	12.3%	0.888±0.008	18.4%
feature-level DF full	0.954±0.004	11.6%	0.900±0.007	17.1%
feature-level DF extended	0.956±0.004	11.2%	0.908±0.007	16.3%
Video				
FF	0.876±0.009	20.4%	0.876±0.009	20.4%
feature-level DF	0.880±0.009	19.7%	0.880±0.009	19.7%
feature-level DF full	0.885±0.008	19.3%	0.885±0.008	19.3%
AV				
FF	0.943±0.005	13.4%	0.914±0.007	16.5%
DF	0.951±0.005	11.7%	0.922±0.006	14.9%
feature-level DF	0.954±0.004	11.6%	0.930±0.006	14.5%
feature-level DF full	0.959±0.004	10.9%	0.935±0.005	13.8%
feature-level DF extended	0.961±0.004	10.5%	0.938±0.005	13.3%
AV weighted				
DF	0.952±0.005	11.6%	0.924±0.006	14.8%
feature-level DF extended	0.960±0.004	10.6%	0.936±0.005	13.6%

Table 1: Area Under the Curve (AUC) and Equal Error Rate (EER) averaged over all 11 speakers and all 7 SNR levels when car noise was added. See the text for details on the features and fusion approaches (FF=Feature Fusion, DF=Decision Fusion).

The results in Table 1 show that the decision fusion outperforms the feature fusion in all cases. This is true when looking on the two modalities individually as well as when combined. Further, for the audio-visual fusion the feature-level decision fusion outperforms the standard decision fusion. Including the features from the feature fusion (feature-level decision fusion full) and also the audio features excluding f_0 or loudness (feature-level decision fusion extended) yields additional improvements. Again this is the case when looking only on the individual modalities or at their fusion[1]. These differences are already present when looking only at the clean signal. In the noisy condition they become significantly larger (compare also Fig.1). The results from the weighted fusion performed identical to the unweighted fusion[2]

6. Discussion

The results show that the assumption of conditional independence on the feature level as expressed in our feature-level decision fusion notably improves the performance. The more features we add the better the performance gets. This is the case for the clean and the noisy condition. The correlations between the features seem not to be problematic. This is particularly visible when we look at the case of further adding the audio features resulting from the feature fusion without f_0 or loudness to the already complete feature set (what we termed extended). Even for this case the performance improves further. For noisy audio this improvement is also statistically significant. The good results of the model assuming class conditional independence suggest that the correlations between the features are distributed evenly across the classes or cancel each other out.

We did not observe a benefit from dynamically weighting the two modalities. We assume that the reason is that in our two class problem not one particular class is selected when the noise increases but rather the probabilities continue to distribute evenly across the two classes. In such a scenario a weighting yields no additional benefit.

Based on the results of the weighted and unweighted fusion we conclude that the class conditional independence assumption is well justified in our scenario. It remains unclear however if this is due to the features we used or the two-class nature of the problem.

[1]We could not perform an experiment equivalence to the extended case for the visual channel as we only use two visual features

[2]The small differences are mainly due to numerical instabilities resulting from the additional power operation.

7. Conclusion

In this paper we investigated different fusion methods in the context of audio-visual word prominence detection. The feature-level decision fusion we propose yielded in all cases the best results. We observed gains of up to 50% compared to the audio only detection when using noisy audio. With an equal error rate of 19.3% also the detection based on the video stream alone showed good performance. From the visual modality we used only rigid head movements and lip movements. It is however known that eye brow movements are also used by some speakers to signal word prominence [45]. As the framework of feature-level decision fusion seems to be able to cope well with features with varying reliability the integration of more visual features is something we want to pursue in the future.

8. Acknowledgments

We want to thank Petra Wagner, Britta Wrede, Heiko Wersing and Andrea Schnall for fruitful discussions. Furthermore, we are very grateful to Rujiao Yan and Samuel Kevin Ngouoko for helping in setting up the visual processing and the forced alignment, respectively as well to Venkatesh Kulkarni for developing the Voicing Detection. 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 and Paschalis Mikias and Merikan Koyun for help in the data preparation. Special thanks go to our subjects for their patience and effort.

9. References

[1] 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.

[2] 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.

[3] A. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning dynamic stream weights for coupled-hmm-based audio-visual speech recognition," *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, vol. 23, no. 5, pp. 863–876, May 2015.

[4] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *Multimedia, IEEE Transactions on*, vol. 2, no. 3, pp. 141–151, 2000.

[5] 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, 2009, pp. 604–609.

[6] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10.

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

[8] 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.

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

[10] 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.

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

[12] M. Swerts, D. Litman, and J. Hirschberg, "Corrections in spoken dialogue systems," in *Sixth Int. Conf. on Spoken Language Proc. (ICSLP)*. Bejing: ISCA, 2000.

[13] D. Litman, J. Hirschberg, and M. Swerts, "Characterizing and predicting corrections in spoken dialogue systems," *Computational linguistics*, vol. 32, no. 3, pp. 417–438, 2006.

[14] F. Tamburini and P. Wagner, "On automatic prominence detection for german," in *Proc. of INTERSPEECH*. Antwerp, Belgium: ISCA, 2007.

[15] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 15, no. 2, pp. 690–701, 2007.

[16] 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.

[17] 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.

[18] B. Streefkerk, "Acoustical correlates of prominence: A design for research," in *Proc. Inst. of Phonetic Sciences of the Univ. of Amsterdam*, vol. 21, 1997, pp. 131–142.

[19] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," 2009, pp. 81–84.

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

[21] M. Heckmann, "Audio-visual word prominence detection from clean and noisy speech," *Submitted to Computer Speech and Language*.

[22] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[23] H. Zhang, "The optimality of naive bayes," in *Proc.Int. Florida Artificial Intelligence Research Society Conference (FLAIRS)*, vol. 1, no. 2, 2004, p. 3.

[24] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust asr," *Speech Communication*, vol. 34, no. 1, pp. 25–40, 2001.

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

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

[27] 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, pp. 2421–2424, 2006.

[28] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2013.

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

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

[31] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in *Proc. INTERSPEECH*, 2005.

[32] 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.

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

[34] R. Stiefelhagen, J. Yang, and A. Waibel, "Tracking eyes and monitoring eye gaze," in *Proceedings of the workshop on perceptual user interfaces*, 1997, pp. 98–100.

[35] M. Nishigaki, S. Rebhan, and N. Einecke, "Vision-based lateral position improvement of RADAR detections," in *Proc. 15th Int. IEEE Conf. on Intell. Transportation Systems (ITSC)*, 2012.

[36] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf on Multimedia*. ACM, 2010, pp. 1459–1462.

[37] Y. Xu and C. X. Xu, "Phonetic realization of focus in English declarative intonation," *Journal of Phonetics*, vol. 33, no. 2, pp. 159–197, 2005.

[38] M. Dohen and H. Lœvenbruck, "Interaction of audition and vision for the perception of prosodic contrastive focus," *Language and speech*, vol. 52, no. 2-3, pp. 177–206, 2009.

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

[40] G.-A. Levow, "Context in multi-lingual tone and pitch accent recognition." in *INTERSPEECH*, 2005, pp. 1809–1812.

[41] 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.

[42] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Machine learning*, vol. 31, pp. 1–38, 2004.

[43] A. Varga and H. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[44] G. Hirsch, "FaNT - Filtering and Noise Adding Tool," Niederrhein University of Applied Sciences, Krefeld, Germany, Tech. Rep., 2005.

[45] E. Krahmer and M. Swerts, "More about brows," *From brows to trust*, pp. 191–216, 2005.