Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 48 (2018) 15–30

www.elsevier.com/locate/csl

Audio-visual word prominence detection from clean and noisy speech☆

Martin Heckmann

Honda Research Institute Europe GmbH, D-63073 Offenbach am Main, Germany

Received 22 December 2015; received in revised form 8 May 2017; accepted 28 September 2017
Available online 5 October 2017

Abstract

In this paper we investigate the audio-visual processing of linguistic prosody, more precisely the detection of word prominence, and examine how the additional visual information can be used to increase the robustness when acoustic background noise is present. We evaluate the detection performance for each modality individually and perform experiments using feature and decision fusion. For the latter we also consider the adaptive fusion with fusion weights adjusted to the current acoustic noise level. Our experiments are based on a corpus with 11 English speakers which contains in addition to the speech signal also videos of the speakers' heads. From the acoustic signal we extract features which are well known to capture word prominence like loudness, fundamental frequency and durational features. The analysis of the visual signal is based on features derived from the speaker's rigid head movements and movements of the speaker's mouth. We capture the rigid head movements by tracking the speaker's nose. Via a two-dimensional Discrete Cosine Transform (DCT) calculated from the mouth region we represent the movements of the speaker's mouth. The results show that the rigid head movements as well as movements inside the mouth region can be used to discriminate prominent from non-prominent words. The audio-only detection yields an Equal Error Rate (EER) averaged over all speakers of 13%. Based only on the visual features we obtain 20% of EER. When we combine the visual and the acoustic features we only see a small improvement compared to the audio-only detection for clean speech. To simulate background noise we added 4 different noise types at varying SNR levels to the acoustic stream. The results indicate that word prominence detection is quite robust against additional background noise. Even at a severe Signal to Noise Ratio (SNR) of −10 dB the EER only rises to 35%. Despite this the audio-visual fusion leads to notable improvements for the detection from noisy speech. We observe relative reductions of the EER of up to 79%.

1. Introduction

When humans communicate they not only listen to what is said but also to how it is said. These prosodic variations play a vital role in human communication (Shriberg, 2005). For spoken dialog systems one situation where the prosodic information is particularly important is after a misunderstanding between the human and the machine (Litman et al., 2006; Levow, 2004). Humans use prosodic cues to highlight a correction following a misunderstanding when talking to another human but also when talking to a machine (Swerts et al., 2000). A distinguishing feature of corrections is that they are frequently hyperarticulated and hence perceived as highly prominent (Litman et al.,

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.
E-mail address: martin.heckmann@honda-ri.de

http://dx.doi.org/10.1016/j.csl.2017.09.002
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

2006). Acoustically, prominence is mainly realized via changes in segment duration and intensity as well as fundamental frequency modifications (Streefkerk, 2002).

At least since the observations of Sumby and Pollack in 1954 that seeing a speaker's face and lip movements improves human speech intelligibility scores in noise (Sumby and Pollack, 1954) it is well acknowledged that the visual modality contains a lot of information relevant to communication. As a result, the field of audio-visual speech recognition emerged which in particular targets the robust recognition in noise (Potamianos et al., 2003; Zhou et al., 2014; Heckmann et al., 2002; Kolossa et al., 2009).

Furthermore, it has been shown that humans are able to use visual information to extract linguistic prosodic cues and in particular word prominence (Graf et al., 2002; Munhall et al., 2004; Beskow et al., 2006; Swerts and Krahmer, 2008; Al Moubayed and Beskow, 2009). Yet so far this information has not been used in systems to extract linguistic prosody. Based on the findings in audio-visual speech recognition one expects that the visual information is particularly beneficial in situations in which the acoustic signal is impaired. Such impairments by additional background noise and reverberations from the room commonly occur when speech processing is applied in real world applications, in particular on hand held devices, with robots and in cars. Due to the typical large distances between the microphones and the speaker's mouth, the impact of these disturbances can be quite high. In the realm of speech recognition this is a topic which has received a lot of attention (see Li et al. (2014) for a current overview). The aforementioned speech impairments in principle also apply to the prosodic speech analysis. Yet much less effort has been spent to cope with these impairments in this context. As far as we are aware of, the impact of speech impairments and viable countermeasures have only be addressed for audio-only emotion classification (Schuller et al., 2007; Eyben et al., 2013; You et al., 2006).

We introduced audio-visual word prominence detection in Heckmann (2012) on a dataset of 3 speakers. In Heckmann (2013) we extended the dataset to 16 speakers and improved the acoustic feature extraction. Next, we included context features, i.e. features spanning across the current segment, and feature contour modeling in Schnall and Heckmann (2014) and Heckmann (2014). In this paper we use the same acoustic feature extraction as in our previous work but introduce an improved visual processing including a correction of the speaker's head tilt. Based on this we evaluated the performance of rigid head movements and movements inside the mouth region compared to the acoustic features. This evaluation showed that there is a large variation in visual detection performance for the different speakers. Nevertheless, overall the visual features contribute a lot of information. Furthermore, we investigated different audio-visual fusion schemes. Here we saw that a fusion on the decision level clearly outperforms a feature fusion. A key aspect of this paper is the evaluation of the prominence detection also from noisy audio signals. We performed this evaluation for different types of noise added to the speech signal at a wide range of Signal to Noise Ratio (SNR) levels. This evaluation also included an evaluation of the importance of two of the main cues to word prominence, fundamental frequency and intensity variations, in varying noise conditions. The results showed that the word prominence detection is quite robust against background noise and that the relative importance of the fundamental frequency and intensity derived features depends on the noise type. A further important part of this evaluation is the assessment of the audio-visual fusion schemes with varying degrees of degradations in the acoustic channel. We will demonstrate that the decision fusion is also superior when the acoustic signal is degraded by noise and that relative improvements of 79% compared to an audio-only detection can be obtained from an audio-visual fusion. However, adaptively weighting the two modalities dependent on the acoustic noise level does not further improve the results.

2. Prior work

Quite a few approaches to detect prosodic word prominence have been proposed in the past (Streefkerk, 2002; Tamburini, 2003; Wang and Narayanan, 2007; Jeon and Liu, 2010; Schillingmann et al., 2011). Some authors rather focus on the detection of pitch accent, one acoustic realization leading to prosodic prominence (Shriberg and Stolcke, 2004; Levow, 2005; Rosenberg and Hirschberg, 2009). All these approaches have in common that they only consider the acoustic modality and assume that the signal is not distorted by noise.

The fusion of acoustic and visual information has received a lot of attention in the past (Atrey et al., 2010). A key question in this domain was to find models to optimally fuse the acoustic and visual modalities (Teissier et al., 1999; Potamianos et al., 2003; Yoshida et al., 2009). Particularly relevant to audio-visual speech recognition are models which are able to cope with the complex temporal relation between auditory and visual cues in speech. Depending on the phoneme and its context, variations in the synchrony of an approx. 200 ms visual lead to an approx. 50ms auditory lead have been observed (Schwartz and Savariaux, 2014). Several fusion models which are able to take these assynchronies explicitly (Abdelaziz et al., 2015; Dupont and Luettin, 2000; Nefian et al., 2002) or implicitly (Heckmann et al., 2002) into account have been proposed. As the perceptual benefits of the additional visual information for humans in the perception of speech degraded by noise had already been observed in Sumby and Pollack (1954) many of these approaches have been evaluated when noise was added to the acoustic channel (Heckmann et al., 2002; Kolossa et al., 2009; Yoshida et al., 2009). During fusion of clean video information with noisy audio it has been observed that a dynamic weighting of the two modalities is often quite beneficial. Based on this different models for the estimation of the stream confidence and adaptive weighting have been developed (Potamianos and Neti, 2000; Heckmann et al., 2002; Abdelaziz et al., 2015).

The impact of additional noise in the speech signal has not yet been investigated in the context of linguistic prosody but only for emotional prosody. Yet, also for emotional prosody only very few papers have been published. The classification of emotions on an acted dataset where white noise and sinusoidal noise was added to the acoustic signal was investigated in You et al. (2006). In Schuller et al. (2007) datasets with acted emotions or recorded when children were interacting with a small robot were used. The authors either used recordings from a distant microphone which already contained reverberations and background noise or added it afterwards to the audio signal. They extended their work in Eyben et al. (2013) by investigating which features are best suited for the emotion classification from noisy speech. The impact of babble noise on the emotion classification was investigated in Kim et al. (2007). All the studies showed that the analysis of emotional prosody is quite robust against noise and reverberations.

Audio-visual classification is more and more frequently applied in emotion classification (Zeng et al., 2009; Valstar et al., 2013). Typically the same fusion methods are deployed in audio-visual emotion classification as in audio-visual speech recognition (Zeng et al., 2009). However, the audio-visual detection of linguistic prosody has not yet been done. So far studies have only investigated the role the visual information plays in human perception of word prominence and how speakers modify their facial and head articulators. Graf et al. concluded that head and face movements are strongly correlated with the prosodic content of the speech signal but at the same time that the code seems to be less clearly defined than for the acoustic channel and also shows a very large variation from speaker to speaker (Graf et al., 2002). Measurements of the visual articulators in the production of different focus conditions showed that prominence is visually signaled via hyperarcticulation by larger mouth opening, lip spreading and lip protrusion (Dohen et al., 2006; Scarborough et al., 2009; Cvejic et al., 2010). Also rigid head movements are used to express word prominence, however they are less consistent and show a larger inter-speaker variability than mouth and jaw movements (Graf et al., 2002; Dohen et al., 2006; Scarborough et al., 2009; Cvejic et al., 2010; Kim et al., 2013). Finally, eyebrow movements are also frequently reported to be a strong cue to prominence (Cavé et al., 1996; Graf et al., 2002; Dohen et al., 2006; Scarborough et al., 2009; Cvejic et al., 2010; Kim et al., 2013). Of all the visual features eyebrow movements seem to be the least consistent and show the largest inter-speaker variations. The question which visual articulators contribute to the perception of word prominence has also been addressed via perceptual experiments where only parts of the head or face were visible to the participants. The results of these experiments are a bit inconclusive. While Swerts and Krahmer (2008) reported that the upper part of the face is more informative for the listeners, Cvejic et al. (2012) stated that it is the lower part. It is reported quite unanimously that the inclusion of the visual information improves the perception either by higher prominence detection scores or faster reaction times (Dohen and Lłevenbruck, 2009; Swerts and Krahmer, 2008; Al Moubayed et al., 2010).

3. Dataset

To stimulate corrections and hence prominent words, we recorded participants interacting via speech in a Wizard of Oz experiment (Heckmann, 2012). In a computer game they instructed the system to move tiles to uncover a cartoon . This game yielded utterances of the form “place green in B one”. Occasionally, a misunderstanding of one word of the sequence was triggered by the experimenter and the corresponding word was highlighted, verbally and visually. Which word of the sequence was misunderstood was determined by a random process, yet the triggering was performed by the experimenter. The triggering was opaque to the participants and the evaluation of the subsequent questionnaires revealed that they believed that it was a true misunderstanding of the machine. Participants were told that after a misunderstanding they should repeat 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 e.g. by 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 participants were recorded (Heckmann, 2013). The audio signal was originally sampled at 48 kHz and later down sampled 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 (Young et al., 1995) on the Grid Corpus (Cooke et al., 2006) followed by a speaker adaptation with a Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995) step with a subsequent Maximum A-Posteriori (MAP) (Gauvain and Lee, 1994) step to perform a forced alignment of the data.

Three human annotators annotated the recorded data with 4 levels of prominence for each word. Unfortunately, there is no generally agreed upon prominence scale. In an effort to make the ratings nevertheless consistent, we manually extracted four words with different levels of prominence from the recorded dataset and made them available to the annotators during their annotation such that they could use it as a reference. In previous experiments a prominence scale with 4 levels had been found to be well adapted (Jensen and Tndering, 2005). 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 approx. 400 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 (Schuller and Batliner, 2013; Rosenberg, 2009). 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. Low-level descriptors

Table 1 gives an overview on all acoustic and visual low-level descriptors we used. In the following we will explain them in more detail.

For the acoustic modality a key feature to word prominence are variations of the perceived signal energy, i.e. loudness variations. We extracted the loudness l by filtering the signal with an 11th order IIR filter as described in Knowledgebase (2016), 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_W, the duration of the word, and D_G the duration of the gap before, respectively after the word. These values were determined from the forced alignment. We also extracted the fundamental frequency f_0 (following Heckmann et al. (2007)), interpolated values in the unvoiced regions via cubic splines and converted the results to semitones. To detect voicing, we used an

Table 1
Acoustic and visual low-level descriptors.

Audio	
l	Loudness in dB
D_W	Duration of the word
D_G	Duration of the gap between words
f_0	Fundamental frequency in semitones
SE	Spectral emphasis, i.e. the difference between overall energy and low-pass energy
Video	
n_x, n_y	Nose x and y position
m_{DCT}	DCT transform calculated from the mouth region

extension of the algorithm described in Kristjansson et al. (2005). 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 $1.5f_0$ (Heldner, 2003).

To extract low-level descriptors from the visual channel, we used the OpenCV library (Bradski, 2000) 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 selected one image. Due to the small movements in the temporal proximity of this image we expect the nose detection to be reliable for this image. From this image we cut out a region around the nose and used it as a template for a correlation-based nose tracking. We then tracked the nose starting from this image by correlating the template with the image and determining the shift between them. As this stable region might not be at the start of the utterance the correlation-based tracking had to be performed forward in time until the end of the utterance as well as backwards in time until the start of the utterance. Once we obtained the nose tracks we also determined the eyes in the image. For doing so we detected the darkest spot in the image where we expected the eyes based on the nose position, a frequently used technique (Stiefelhagen et al., 1997). 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 Nishigaki et al. (2012). 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 10,000 coefficients per image we selected the 20 with the lowest spatial frequencies.

4.2. 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). Next, we determined word boundaries via a forced alignment. As functionals we then extracted the mean, max, min, spread (max-min) and variance along the word.

To capture all the information in the feature and to be more tolerant against noise a more holistic representation based on contours is promising. Different methods to robustly model contours with a few descriptive coefficients have been proposed in the literature. We could previously show that adding such holistic contour models to the functionals improves the performance (Heckmann, 2014). As functional Principal Component Analysis (functional PCA) (Arias et al., 2013), the method which performed best in in our previous experiments, requires a learning step of the basis functions, we decided to use the DCT instead, a frequently used and computationally very simple method (Eyben et al., 2010), which yielded almost identical results to the functional PCA (Heckmann, 2014). The DCT only requires a summation and multiplication with a fixed set of coefficients

$$y(k) = \sum_{n=1}^{N} x(n) \cos\left[\frac{\pi}{2}\left(n - \frac{1}{2}\right)(k-1)\right], \qquad (1)$$

where N is the length of the segment and $\mathbf{x}$ an acoustic or visual low-level descriptor after the previously mentioned normalization. Effectively, the DCT transforms the contour into a frequency representation. By retaining from the set of DCT coefficients $y(1), \ldots, y(N)$ only the first K coefficients we obtain a representation which solely captures the low frequency variations in the signal. We used different values K_A and K_V for the acoustic and visual modality. 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. In case of the features derived from the mouth region this means that we first calculated a two-dimensional DCT along the image dimensions to capture the intensity variations in one single image and then a one-dimensional DCT along the time dimension to capture variations of the previously calculated two-dimensional DCT coefficients over time.

4.3. Context features

Marking the focus of a word in an utterance, rendering it prominent, also has an influence on the neighboring words: the word in focus is hyperarticulated and the surrounding words are hypoarticulated (Xu and Xu, 2005; Dohen and Lłevenbruck, 2009). It has been shown previously that taking this context information into account is very effective for the detection of word prominence (Schnall and Heckmann, 2014; Levow, 2005; Rosenberg and Hirschberg, 2009). Therefore, we also applied 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 Schnall and Heckmann (2014) for details):

$$\text{Where} \quad \mathbf{o}_{\text{Context}} = [\mathbf{o}_{m-1}, \mathbf{o}_m, \mathbf{o}_{m+1}] \in \mathbb{R}^{l_{\text{Context}}}, \quad l_{\text{Context}} = 3l_F, \tag{2}$$

with $\mathbf{o}_m$ the functionals of word m and l_F the dimensionality of the vector of functionals of a single word.

In summary, our feature extraction process consists of

- Extraction of low-level descriptors from the acoustic and visual channel,
- Calculation of derivatives from the descriptors,
- Segmentation of the utterance into words,
- Normalization to the utterance mean,
- Contour representation and functional calculation,
  - Modeling of the contour on a word level via a DCT transform,
  - Calculation of functionals on the word level,
- Concatination of feature vectors to context features.

5. Audio-visual fusion

For the fusion of the audio and video modality we evaluated fusion models which are commonly used in audio-visual speech recognition (Potamianos et al., 2003): feature and decision fusion. In case of the decision fusion we also investigated if, similar to audio-visual speech recognition, a weighting of the two modalities dependent on the current noise scenario yields better results. As the word prominence detection operates on the word level we do not expect a benefit from modeling the asynchronies in the audio-visual fusion. Hence, we only investigated models which assume synchrony between the two modalities.

5.1. Feature fusion

We implemented the feature fusion, also called feature concatenation, as Potamianos et al. (2003):

$$\mathbf{o}_{AV} = [\mathbf{o}_A, \mathbf{o}_V] \in \mathbb{R}^{l_{AV}}, \tag{3}$$

where $\mathbf{o}_A$ and $\mathbf{o}_V$ are the functionals derived from the acoustic and visual modality, respectively. The dimensionality of the corresponding vectors is l_A and l_V yielding $l_{AV} = l_A + l_V$. Hence, we concatenated the feature vectors of the two modalities to form a larger feature vector.

5.2. Decision fusion

In decision fusion the two modalities are classified individually and then the decisions of the individual classifiers are fused. As decisions we used the posterior probabilites $P(C_i|\mathbf{o})$ of functional $\mathbf{o}$ belonging to class C_i, in our case prominent or non-prominent. This posterior probability was provided by the SVMs which we used for classification. While doing so we assumed class-conditional independence between the two modalities

$$P(\mathbf{o}_A, \mathbf{o}_V|C_i) = P(\mathbf{o}_A|C_i)P(\mathbf{o}_V|C_i). \tag{4}$$

Using Bayes formula one derives at (Heckmann et al. 2002)

$$P(C_i|\mathbf{o}_A,\mathbf{o}_V)=\frac{P(C_i|\mathbf{o}_A)P(C_i|\mathbf{o}_V)}{P(C_i)}\eta(\mathbf{o}_A,\mathbf{o}_V), \tag{5}$$

where $P(C_i)$ represents the prior probability of class C_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. In the machine learning community this is also called a naïve Bayesian model and also when the underlying assumption of class-conditional independence is not fully met it usually yields good results (Bishop, 2006)[1]. In the case of audio-visual speech recognition it has been shown previously that the assumption of class-conditional independence can be made (Movellan and McClelland, 2001) and it is very frequently used (Potamianos et al., 2003; Heckmann et al., 2002). As word prominence is to a large extent expressed via a hyperaticulation of the speech articulators it can be expected that these assumptions translate from audio-visual speech recognition to audio-visual word prominence detection.

5.3. Weighted decision fusion

To investigate if a weighting of the two modalities during fusion, as is frequently done in audio-visual speech recognition, is also beneficial for our task, we also implemented a common weighting scheme (Potamianos et al., 2003; Heckmann et al., 2002):

$$P(C_i|\mathbf{o}_A,\mathbf{o}_V)=\frac{P(C_i|\mathbf{o}_A)^{\lambda}P(C_i|\mathbf{o}_V)^{1-\lambda}}{P(C_i)}\eta(\mathbf{o}_A,\mathbf{o}_V), \tag{6}$$

where λ is the weighting parameter varying in the range [0, 1].

6. Results

In this section we will first detail our experimental setup and then present the results for the acoustic and visual modality separately and when combined in clean conditions. Following this we will present results for audio-only detection with additional background noise and corresponding audio-visual results.

6.1. Experimental setup

In previous experiments we compared the performance of a Gaussian Mixture Model (GMM), a Support Vector Machine (SVM), a Conditional Random Field (CRF) and a Deep Neural Network (DNN) to discriminate prominent from non-prominent words using the dataset described in Section 3 (Schnall and Heckmann (0000, 2014, 2016)). For the DNN we used a fully connected feedforward network with 1 to 5 hidden layers (Schnall and Heckmann (0000, 2016)). The results showed that the CRF had a slight lead over SVM and DNN. The latter two performed practically identical. All three methods clearly outperformed the GMM. Due to their very similar performance at a much lower computational complexity we decided to use SVMs for the subsequent experiments. We implemented the SVMs with a Radial Basis Function Kernel using LibSVM (Chang and Lin, 2011). For each feature combination we performed 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. Prior to the grid search, we normalized the features to the range $[-1,\ldots,1]$. With the found optimal parameters we trained an SVM on 75% of the data and tested on the remaining 25%. We repeated this step 30 times to perform a 30-fold cross-validation. To establish the 30 sets, we applied a sampling with replacement strategy where we set the number of elements from the prominent and non-prominent class corresponding to their respective frequency in the dataset. We performed this process individually for each speaker. Consequently, all results are speaker-dependent. The decision level fusion described in Section 5.2 is based on posterior probabilities. Per se an SVM only returns class membership and the distance to the closest support vector. To obtain the posterior probabilities from the SVM we enabled the corresponding option in the LibSVM which calculates an estimate of the posterior probability based on the distance to the support vectors.

[1] Please note that it is commonly assumed in a naïve Bayesian model that all features are conditionally independent to each other whereas we only assume the features of the two modalities to be conditionally independent.

As the two classes in our dataset are very unbalanced (there are roughly 10 times more non-prominent than prominent words), we use Receiver Operating Characteristics (ROCs) to measure the performance of the SVM. The ROC allows one to visually judge the power of a binary classifier for all settings of the decision threshold by plotting the true positive rate against the false positive rate. When calculating the area of the curve covered by the ROC, the so called Area Under the Curve (AUC), a measure of aggregated classification performance can be obtained. An alternative is the Equal Error Rate (EER). However, the EER is only a point measure in the ROC at the position where the false negative rate (1- true positive rate) and the false positive rate are equal. The AUC has values between 1 (the optimal classifier) and 0.5 (a random decision). As already very small variations of the AUC can represent large differences in classification accuracy for a given threshold the AUC values do not allow for an intuitive interpretation. Yet several statistical tests have been proposed to calculate confidence intervals for the AUC values (Qin and Hotilovac, 2008). Hence, they are well suited to assess the statistical significance of results. On the other hand, EERs are much easier to interpret as they are certainly more familiar to most readers. In our experiments we saw that AUC and EER were highly correlated. For these reasons we will use both measures in the following experiments. When we focus on the statistical validity we will rather report the AUC. In cases when we think that the easy interpretability is important we will rather use EER.

We established the ROCs as proposed in Fawcett (2004) by pooling the results of all cross-validations and all speakers and calculated the AUC and the EER.

To investigate the impact of acoustic noise, we added to the clean audio signal white, babble, car and factory noise taken from the Noisex database (Varga and Steeneken, 1993). We adjusted the SNR levels using the tool Fant (Hirsch, 2005) in a range of −10 dB to 15 dB in 5 dB steps. Via a grid search for each scenario, i.e. noise type and SNR level, and averaging over all speakers we determined the optimal setting of the weighting parameter λ in the weighted decision fusion.

6.2. Audio-only detection

For the word prominence detection based only on acoustic information we used three acoustic feature combinations: features, i.e. functionals, derived from all acoustic low-level descriptors described in Section 4.1, all but f_0 or all but loudness. Averaged over all speakers we obtain an EER of 12.9% using all acoustic low-level descriptors[2]. When we remove either f_0 or loudness we see a clear rise in error rate, most notable when we remove f_0 (compare also Table 2). We have chosen f_0 and loudness as they are the most important features for prominence detection[3].

6.3. Visual detection

The visual detection is based on two different types of visual features. One captures the dynamics of the rigid head movements based on the nose movements. The other is related to the non-rigid movements of the speaker's mouth, e.g. lip movements. To extract these we used the DCT calculated on the mouth region.

Table 2 compares the results averaged over all speakers for these two feature types and their combination. As can be seen the mouth related features perform much better than the nose features. Nevertheless, already based on the nose information alone we obtain AUCs and EERs well above chance level. This is depicted in more detail in Fig. 1a. Here the results for each individual speaker are given. As can be seen there is a large variation from speaker to speaker. When considering only the nose movements we obtain an EER of 45.2% for the worst speaker compared to 16.5% for the best. The combination of the two visual feature types improves the performance and also reduces the variance between the speakers. Hereby the mouth features alone perform almost as good as the combination of nose and mouth features. When applying a t-test at an α value of 0.05 we saw no statistically significant difference in performance between either mouth features alone or combined with the nose features. With an EER of 20.4% the visual features perform worse than the audio features (12.9%) but clearly contain a lot of information.

[2] Bear in mind that for this two class problem guessing would yield an EER of 50%.

[3] This is together with duration. Yet as the temporal alignment did not change with the noise level so did the duration.

Table 2
Area Under the Curve (AUC) with 95% confidence intervals[a] and Equal Error Rate (EER) averaged over all 11 speakers for clean audio and varying feature usage.

Features	AUC	EER (%)
Audio		
All features	0.943 ± 0.005	12.9
w/o f_0	0.919 ± 0.006	15.8
w/o loudness	0.927 ± 0.006	14.5
Video		
Nose	0.770 ± 0.012	29.0
Mouth	0.869 ± 0.009	20.9
All features	0.876 ± 0.009	20.4
Audio-visual		
Feature fusion	0.943 ± 0.005	13.4
Decision fusion	0.951 ± 0.005	11.7
Weighted Decision fusion	0.952 ± 0.005	11.6

[a] We used the Mann-Whitney estimator (Qin and Hotilovac, 2008) in the implementation of Lau (2014) to calculate the confidence intervals.

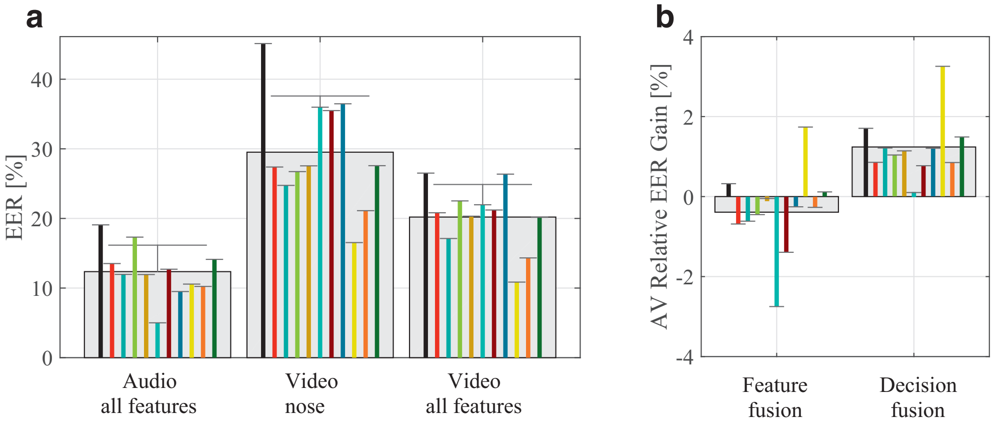

Fig. 1. In (a) the Equal Error Rates (EER) in % for each individual speaker for different feature combinations are shown (thin colored bars). The mean over all speakers is displayed in gray in the background. The corresponding standard deviation over all speakers is indicated via a thin horizontal line on top. In (b) the reduction of the EER in % from audio-visual fusion for each individual speaker is shown. All results are for clean audio.

6.4. Audio-visual detection

When combining visual and acoustic information we consider the two fusion schemes feature fusion and decision fusion, outlined in Section 5. As can be seen from Table 2 the feature fusion does not improve the performance compared to an audio-only detection when we look at the AUC. It slightly degrades performance when considering the EER. On the other hand, the decision fusion leads to a reduction of the errors. The AUC increases from 0.943 to 0.951 (and the EER falls from 12.9% to 11.7%). This difference is also statistically significant. To further investigate this, in Fig. 1b the reduction of the EER from audio-visual fusion for each individual speaker is shown. As can be seen, for the feature fusion the gain is high for one speaker but negative for almost all other speakers. On the other hand, in the case of decision fusion we see at least small to moderate gains for all speakers and a substantial gain of 3.3% absolute for one speaker. The results in Table 2 show that the additional weighting in the weighted decision fusion did not further improve the results.

Table 3
Statistics of the Equal Error Rates for each individual cross validation of all speakers for clean audio and varying feature usage.

Features	⌀(%)	σ(%)	min (%)	max (%)
Audio				
all features	12.9	4.3	2.4	26.0
Video				
all features	20.4	5.6	4.8	35.3
Audio-visual				
feature fusion	13.4	4.5	3.0	26.5
Decision fusion	11.7	4.3	2.5	23.6

Finally, Table 3 shows some statistics on the individual cross validations of all speakers. From the values calculated from all cross validations we see that the standard deviation is in general quite high but it is particularly high for the video-only detection. The feature fusion is not able to deal with these large variations in the visual channel as well as the decision fusion. When using the feature fusion not only the mean value but also the standard deviation increases compared to the audio-only detection. With the decision level fusion the mean decreases and the standard deviation remains unchanged. This demonstrates that the decision fusion is much better suited to our task.

6.5. *Audio-only detection from noisy audio*

Next, we want to investigate the performance of the audio-only word prominence detection when the audio signal is corrupted by noise. For the noisy signals we made two assumptions:

1. The noise type and SNR level are known during training time. I.e. we trained and tested the SVM for identical noise type and SNR,
2. The temporal alignment from the forced Viterbi is not affected by the additional noise.

We made assumption (1) as on the one hand training on clean data and testing on noisy data is not very realistic. As an alternative training with a larger set of noise types and SNR levels can be performed. Yet the selection of the noise conditions also strongly biases the results. Hence we think that our selection of training and testing in the same noise condition will give optimistic but still realistic results. Regarding assumption (2), we could show previously that the exact temporal alignment is not critical (Heckmann et al., 2014). Yet assuming that the alignment is not affected by the noise will give again rather optimistic results. However, using the output of a speech recognition system in the same noise conditions will on the other hand render the results dependent on the performance of this particular speech recognition system in noise. In short, in our experiments we mainly evaluated the effect of the noise on the prosodic features and subsequent functionals and hence will obtain rather optimistic results compared to an application in a real system.

Fig. 2a–d show the results for varying SNR levels of added noise averaged over all speakers depicted as AUC. From the plots one can see that the impact of the noise on the acoustic detection strongly depends on the noise type and is in general not drastic. We see the worst performance for car noise at an SNR of −10 dB with an AUC of 0.684 and an EER of 35.1% (bear in mind that chance level would be at an AUC of 0.5 and an EER of 50%). On the other hand, for white noise, even at −10 dB, performance is still at an AUC of 0.867 and an EER of 16.6%.

In Table 4 the results for a given noise type averaged over all SNR levels are shown. These results show again that the impact on the detection performance varies a lot with the type of noise.

In the same way as for clean audio we also performed experiments for noisy audio where we removed either f_0 or loudness from the set of acoustic low-level descriptors. As can be seen from Table 4 and Fig. 2a–d the performance decreases if we remove either f_0 or loudness from the set of acoustic features. The results also show that it depends on the noise type which feature is more important. For babble noise loudness is the more important feature (compare Fig. 2b), whereas for car noise it is f_0 (compare Fig. 2c). Removing either feature has similar effects on the detection with white or factory noise present. Overall, the degradation from removing a feature is notable but not drastic. We

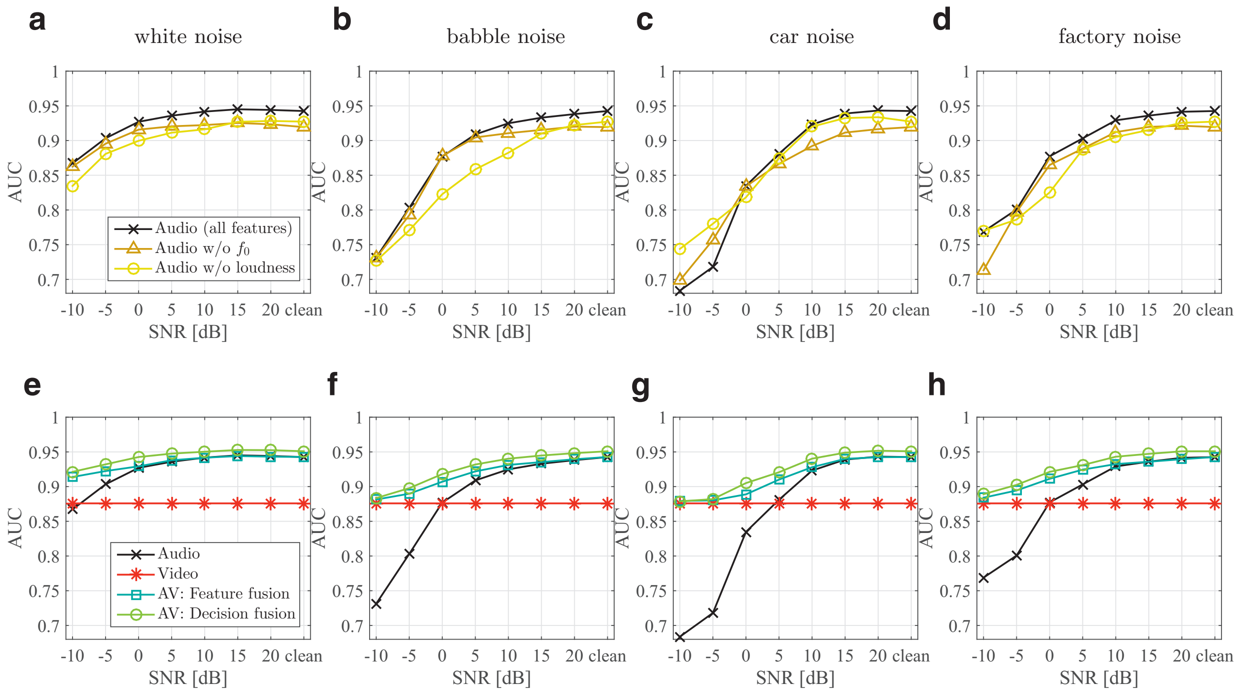

Fig. 2. Area Under the Curve (AUC) averaged over all 11 speakers with varying SNR levels and varying feature usage. In a−d audio-only results with all acoustic features or when excluding one are shown. The comparison between audio-only, video-only and audio-visual detection is shown in e−h.

see the largest effect with babble noise added at an SNR of 0 dB with a relative increase of EER of 27% when we remove the loudness feature.

6.6. Audio-visual detection from noisy audio

Finally, we investigate how the performance in noise can be improved by incorporating the visual information. Also in this case we investigated feature and decision fusion. As can be seen from Fig. 2e−h and Table 4 the additional visual information substantially improves the performance. The observable degradation with increasing noise level is much slower for the audio-visual case. Furthermore, the audio-visual performance never drops below the visual only performance. In most cases there is a synergy effect from combining the two modalities, i.e. the combined results are better then either result by itself. Again the decision fusion performs better than the feature fusion.

Table 4
Area Under the Curve (AUC) with 95% confidence intervals and Equal Error Rate (EER) for a given noise type averaged over all 11 speakers and all 8 SNR levels. See the text for the meaning of the feature names and modeling abbreviations.

Features	White	Babble	Car	Factory
Audio				
All features	0.926 ± 0.006 (14.4%)	0.882 ± 0.008 (18.7%)	0.858 ± 0.008 (20.5%)	0.887 ± 0.008 (18.3%)
w/o f_0	0.911 ± 0.007 (16.5%)	0.871 ± 0.008 (19.9%)	0.849 ± 0.009 (21.6%)	0.867 ± 0.008 (20.4%)
w/o loudness	0.903 ± 0.007 (16.7%)	0.853 ± 0.009 (21.8%)	0.866 ± 0.008 (20.2%)	0.868 ± 0.009 (19.8%)
Video				
All features	0.876 ± 0.009 (20.4%)	0.876 ± 0.009 (20.4%)	0.876 ± 0.009 (20.4%)	0.876 ± 0.009 (20.4%)
Audio-visual				
Feature fusion	0.934 ± 0.006 (14.1%)	0.919 ± 0.007 (16.0%)	0.914 ± 0.007 (16.5%)	0.921 ± 0.007 (15.7%)
Decision fusion	0.944 ± 0.005 (12.5%)	0.927 ± 0.006 (14.6%)	0.922 ± 0.006 (14.9%)	0.929 ± 0.006 (14.1%)
Weighted Decision fusion	0.943 ± 0.005 (12.6%)	0.928 ± 0.006 (14.5%)	0.924 ± 0.006 (14.8%)	0.930 ± 0.006 (14.0%)

In particular, for decision fusion we see synergy effects from the combination of both modalities for all cases tested. The largest reduction of EER we observe is from 35.1% to 19.6% for car noise at an SNR of -10 dB with decision fusion. This corresponds to a relative improvement of 79%. Setting weights depending on the current noise scenario on the audio and video modality during fusion did not further improve the results.

7. Discussion

First of all, with our experiments we could show that the visual channel contains a lot of information for the detection of word prominence. Considering only the rigid head movements, captured by the movements of the nose, we see EERs above chance level for all speakers. At the same time we see a large inter-speaker variation in EER from 16.5% up to 45.2% with an average of 29.0%. Relatively larger inter-speaker variations in the production of visual prosody compared to acoustic prosody have been observed before (Graf et al., 2002; Dohen and Łœvenbruck, 2009). The mouth movements seem to be more consistent and more informative. They yield EERs in the range from 11.1% to 28.7% with an average of 20.9%. Previous studies did not come to a clear consensus on which facial areas convey most information on word prominence. In Cvejic et al. (2012) listeners had to discriminate broad and narrow focus from the presentation of different parts of the face of a speaker. The results showed that the lower part of the face was more informative for the listeners. In contrast to this, in a very similar experiment in Swerts and Krahmer (2008) it was reported that the upper part of the face was more informative for the listeners. This difference might be explained by the large inter-speaker variations in expressing visual prosody (Cvejic et al., 2012). Furthermore, it can be expected that the mouth region yields more consistent information on prominence as it is closely related to relevant acoustic realizations. Important cues for prominence, like duration and amplitude, are easily visible from the mouth region (Kochanski et al., 2005). In contrast to this, fundamental frequency is more linked to rigid head and eyebrow movements (Cavé et al., 1996). As we did not extract the eyebrow movements this can also explain the clear advantage of the mouth features compared to the features related to the upper head, i.e. rigid head movements. The combination of the mouth movement features with the less reliable features derived from the rigid head movements did only lead to a small and statistically not significant improvement to 20.4%. Overall, this means that the EER we obtain from the purely visual detection is only approx. 60%, in relative terms, higher than that of the audio-only detection. In our data we saw that the speakers rotated their heads around all three axis (i.e. nodding in the sagittal plane, turning in the transverse plane and tilting in the frontal plane) at varying degrees while speaking. With our recording setup and the processing steps laid out in Section 4.1 we were only able to compensate for the tilting of the head. Consequently, in some cases we could not extract the visual features correctly. From visual inspection we observed a relation between the head rotations and the visual detection performance. The speaker of whom we obtained the worst results rotated his head the most whilst the speaker of whom we obtained the best results showed very little head rotations. Hence, we expect that substantially more information can be extracted from the visual channel once the features can be extracted independently of these head movements. For our current data this would require 3D models of the speakers' heads. For future experiments we think it is promising to use 3D or 2.5D recordings, e.g. with an active sensor as the Microsoft Kinect. Furthermore, our visual features do not capture all available visual information, e.g. eyebrow movements and frowning. It can be expected that including also these features will further enhance the detection performance.

Despite the impairments of the visual features via the head rotations, we could show that adding the visual information to the acoustic information improves the detection of word prominence, although only to a small extent and only for the decision fusion. One could expect that the feature fusion, which is able to capture all the correlations in the different modalities, would yield better results than the decision fusion, which makes in our case the assumption of class-conditional independence. We suppose one of the reasons for the superior performance of the decision fusion is the at times varied quality of the visual features. Depending on the subset in the different cross validations the visual features might look quite different in the training and test set. This is also indicated by the large variance of the individual cross validation results for visual only detection (compare Table 3). The decision level fusion is known to be able to compensate much better for such mismatches between training and testing conditions in a single modality as the feature fusion (Heckmann et al., 2002). The increase in the variance of the individual cross validation results for the feature fusion compared to the audio-only detection and an unchanged variance for the decision fusion also reflects this.

In a next step we investigated the impact of additional background noise in the acoustic channel on the audio-only and audio-visual detection. The impact of the noise on the detection performance depends not only largely on the noise level but also on the noise type. Yet, as we saw the detection of word prominence is in general very tolerant against background noise. Even at a severe SNR of −10 dB the error rates only reach 26% compared to 13% in clean conditions. In the context of emotion recognition from noisy speech it has already been stated that the prosodic analysis is more robust against noise than speech recognition (Schuller et al., 2007). This can be further substantiated by comparing the results of word prominence detection to those of a speech recognition experiment in the same acoustic environment. In Heckmann et al. (2011) the recognition of numbers from the clean signal yielded an error rate of 1%[4] while in our experiment the word prominence detection results in an error rate of 13%. When factory noise is added at −5 dB we observe an error rate of 82% for the speech recognition compared to 26% for the word prominence detection. To counterbalance the large differences in task complexity, i.e. discrimination of 11 numbers versus a two class problem, one can also calculate the Relative Improvement Over Chance (RIOC) index[5] (Farrington and Loeber, 1989):

$$\mathrm{RIOC} = \frac{\mathrm{Accuracy} - \mathrm{Chance\ Rate}}{1 - \mathrm{Chance\ Rate}}, \tag{7}$$

The RIOC can take a minimum value of 0 and a maximum of 1. It yields values of 0.99 and 0.74 for clean audio in speech recognition and prominence detection, respectively. In noisy audio with factory noise at −5 dB values drop to 0.14 and 0.47. Hence, also in terms of RIOC prominence detection degrades much slower than speech recognition. From this we conclude that prosodic processing is substantially more robust against additional background noise than speech recognition.

An analysis of the contribution of fundamental frequency and energy variations, the main features to capture word prominence, in clean and noisy conditions revealed that their contribution dependents on the noise characteristics. In the speech like babble noise the fundamental frequency feature is clearly impaired and does not add information at SNR levels below 10 dB. On the other hand in the instationary car noise the situation is reversed and the loudness feature does not contribute for medium SNR levels. At SNR levels below 0 dB both features seem to be impaired to an extent that they yield inconsistent information such that the detection without either feature yields better results than when combining all features.

As the visual channel already improves the performance in clean conditions we expected an even larger contribution in cases in which the speech signal is corrupted by background noise. This is indeed what our results show. With increasing noise level the improvements of the audio-visual detection compared to the audio-only detection get larger. The results also showed that the acoustic and the visual channel contain complementary information as the combination of both is always better than either one taken alone. Here again, the decision fusion is better able to take advantage of this. The results of the feature fusion are always inferior. Compared to the audio-only detection we see typical SNR gains of 5 dB to 10 dB and relative improvements of up to 79% when using the decision fusion. In contrast to results published for audio-visual speech recognition, 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 as during the fusion the unreliable classifier does not impair the results of the more reliable classifier. The weighting is particularly beneficial if the unreliable classifier selects one class with a high confidence such that this selection will then also dominate the fusion result. The matched training we performed underestimates the effect of the noise on the audio channel. Hence, we expect notably larger improvements for more realistic settings where the noise is not known before. On the other hand, the extraction of the visual features is—in less controlled settings—more challenging. This will most likely lead to inferior visual features and, as a probable consequence, to reduced gains from the audio-visual fusion.

[4] To make it comparable to the word prominence detection task we only counted the confusion errors, i.e. the insertions and deletions occurring in a continuous recognition task were not taken into account.

[5] We assumed a maximal accuracy of 1 and chance rates of 1/11 and 1/2, respectively.

8. Conclusion

This paper introduced audio-visual word prominence detection. We analyzed the detection performance from clean and noisy speech. The analysis of the purely visual detection showed that for our dataset the rigid head movements as well as the movements of the mouth contain substantial information on the prominence of a word. When combining these two visual features we achieve an EER of approx. 20%. We compared feature and decision fusion for the audio-visual fusion and saw that decision fusion shows better performance in clean and noisy conditions. For clean conditions we saw a relative improvement of approx. 10% and for noisy conditions of up to approx. 80%. The analysis of the audio-only detection in noise showed that the relative contribution of the f_0 and loudness feature depends on the noise type. Overall we observed that word prominence detection is more robust against background noise than speech recognition.

Our recording setup and the image processing algorithms we devised were not able to cope with the full range of head motions our participants showed during recording. To overcome these limitations more work has to be spent in the extraction of visual features. We expect that this can lead to much better performance for the visual modality. In general, we saw large variations from speaker to speaker. This was the case for the acoustic modality but even more so for the visual modality. The recording of more speakers is necessary to be able to better understand these variations and to build models which can cope with them. For the acoustic modality we currently develop a speaker adaptation method which allows counterbalancing these speaker specific variations (Schnall and Heckmann (0000, 2016)). However, similar approaches are needed for the visual modality. These improvements will form an ideal basis to not only investigate impairments in the acoustic but also in the visual modality. This will then allow exploring the dynamic fusion of the two modalities dependent on their reliability.

Acknowledgments

I want to thank Andrea Schnall, Petra Wagner, Britta Wrede and Heiko Wersing for fruitful discussions. Many additional thanks to Andrea Schnall for comments on the manuscript and the contribution of the context features and other aspects of the processing chain. 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 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 my participants for their patience and effort. I am also very thankful to the anonymous reviewers which gave very helpful comments to improve the paper.

References

Abdelaziz, A., Zeiler, S., Kolossa, D., 2015. Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process 23 (5), 863–876.

Al Moubayed, S., Beskow, J., 2009. Effects of visual prominence cues on speech intelligibility. In: Proceedings of International Conference on Auditory Visual Speech Processing (AVSP). ISCA, Norwich, UK.

Al Moubayed, S., Beskow, J., Granström, B., 2010. Auditory visual prominence: from intelligibility to behavior. J. Multimodal User Interfaces 4 (3), 299–311.

Arias, J.P., Busso, C., Yoma, N.B., 2013. Energy and F0 contour modeling with functional data analysis for emotional speech detection. In: Proceedings of INTERSPEECH. ISCA, Lyon, Fance.

Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S., 2010. Multimodal fusion for multimedia analysis: a survey. Multimed. Syst. 16 (6), 345–379.

Beskow, J., Granström, B., House, D., 2006. Visual correlates to prominence in several expressive modes. In: Proceedings of INTERSPEECH. ISCA, Pittsburgh, PA, USA.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning, 4th Edition. Springer, New York, USA.

Bradski, G., 2000. The openCV library. Dr. Dobb's J. Softw. Tools 11 (25), 120–126.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., Espesser, R., 1996. About the relationship between eyebrow movements and FO variations. In: Proceedings of International Conference Spoken Language Processing (ICSLP). IEEE, Philadelphia, PA, USA.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (27), 1–27. 27, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. 120 (5), 2421–2424.

Cvejic, E., Kim, J., Davis, C., 2012. Recognizing prosody across modalities, face areas and speakers: examining perceivers sensitivity to variable realizations of visual prosody. Cognition 122 (3), 442–453.

Cvejic, E., Kim, J., Davis, C., Gibert, G., 2010. Prosody for the eyes: quantifying visual prosody using guided principal component analysis. In: Proceedings of INTERSPEECH. ISCA, Makuhari, Japan.

Dohen, M., Lłevenbruck, H., 2009. Interaction of audition and vision for the perception of prosodic contrastive focus. Lang. Speech 52 (2−3), 177–206.

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

Dupont, S., Luettin, J., 2000. Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimed. 2 (3), 141–151.

Eyben, F., Weninger, F., Schuller, B., 2013. Affect recognition in real-life acoustic conditions-a new perspective on feature selection. In: Proceedings of INTERSPEECH. ISCA, Lyon, France.

Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of Int. Conf on Multimedia. ACM, Florence, Italy.

Farrington, D.P., Loeber, R., 1989. Relative improvement over chance (RIOC) and PHI as measures of predictive efficiency and strength of association in 2 × 2 tables. J. Quant. Criminol. 5 (3), 201–213.

Fawcett, T., 2004. ROC graphs: notes and practical considerations for researchers. Mach. Learn. 31 (1), 1–38.

Gauvain, J., Lee, C., 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. IEEE Trans. Speech, Audio Process 2 (2), 291–298.

Graf, H., Cosatto, E., Strom, V., Huang, F., 2002. Visual prosody: facial movements accompanying speech. In: Proceedings of International Conference on Automatic Face and Gesture Recognition. IEEE, Washington, D.C., USA.

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

Heckmann, M., 2013. Inter-speaker variability in audio-visual classification of word prominence. In: Proceedings of INTERSPEECH. ISCA, Lyon, France.

Heckmann, M., 2014. Steps towards more natural human-machine interaction via audio-visual word prominence detection. In: Bck, R., Bonin, F., Campbell, N., Poppe, R. (Eds.), Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction. Lecture Notes on Artificial Intelligence. Springer, pp. 15–24.

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

Heckmann, M., Domont, X., Joublin, F., Goerick, C., 2011. A hierarchical framework for spectro-temporal feature extraction. Speech Commun. 53 (5), 736–752.

Heckmann, M., Joublin, F., Goerick, C., 2007. Combining rate and place information for robust pitch extraction. In: Proceedings of INTERSPEECH. ISCA, Antwerp, Belgium.

Heckmann, M., Mikias, P., Kolossa, D., 2014. The impact of word alignment accuracy on audio-visual word prominence detection. In: Proceedings of the 11th ITG Conference on Speech Communication. ITG, Erlangen, Germany.

Heldner, M., 2003. On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in swedish. J. Phon. 31 (1), 39–62.

Hirsch, G., 2005. FaNT - Filtering and Noise Adding Tool. Technical Report. Niederrhein University of Applied Sciences, Krefeld, Germany.

Jensen, C., Tndering, J., 2005. Choosing a scale for measuring perceived prominence. In: Proceedings of INTERSPEECH. ISCA, Lisbon, Portugal.

Jeon, J., Liu, Y., 2010. Syllable-level prominence detection with acoustic evidence. In: Proceedings of INTERSPEECH. ISCA, Makuhari, Japan.

Kim, E.H., Hyun, K.H., Kim, S.H., Kwak, Y.K., 2007. Speech emotion recognition using Eigen-FFT in clean and noisy environments. In: Proceedings of IEEE International Symposium Robot and Human Interactive Communication (RO-MAN). IEEE, Jeju, Korea.

Kim, J., Cvejic, E., Davis, C., 2013. Tracking eyebrows and head gestures associated with spoken prosody. Speech Commun. 57, 317–330.

Knowledgebase, H., 2016. http://wiki.hydrogenaud.io/index.php?title=Replay_Gain.

Kochanski, G., Grabe, E., Coleman, J., Rosner, B., 2005. Loudness predicts prominence: fundamental frequency lends little. J. Acoust. Soc. Am. 118 (2), 1038–1054.

Kolossa, D., Zeiler, S., Vorwerk, A., Orglmeister, R., 2009. Audiovisual speech recognition with missing or unreliable data. In: Proceedings of International Conference on Auditory Visual Speech Process. (AVSP). ISCA, Norwich, United Kingdom.

Kristjansson, T., Deligne, S., Olsen, P., 2005. Voicing features for robust speech detection. In: Proceedings of INTERSPEECH. ISCA, Lisbon, Portugal.

Lau, B., 2014. MatlabAUC toolbox. https://github.com/brian-lau/MatlabAUC.

Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Lang. 9 (2), 171–185.

Levow, G., 2004. Identifying local corrections in human-computer dialogue. In: Proceedings of INTERSPEECH. ISCA, Jeju, Korea.

Levow, G.-A., 2005. Context in multi-lingual tone and pitch accent recognition. In: Proceedings of INTERSPEECH. ISCA, Lisbon, Portugal.

Li, J., Deng, L., Gong, Y., Haeb-Umbach, R., 2014. An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process 22 (4), 745–777.

Litman, D., Hirschberg, J., Swerts, M., 2006. Characterizing and predicting corrections in spoken dialogue systems. Comput. Linguist. 32 (3), 417–438.

Movellan, J.R., McClelland, J.L., 2001. The Morton-Massaro law of information integration: implications for models of perception. Psychol. Rev. 108 (1), 113.

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

Nefian, A.V., Liang, L., Pi, X., Liu, X., Murphy, K., 2002. Dynamic Bayesian networks for audio-visual speech recognition. EURASIP J. Appl. Signal Process 2002 (1), 1274–1288.

Nishigaki, M., Rebhan, S., Einecke, N., 2012. Vision-based lateral position improvement of RADAR detections. In: Proceedings of IEEE Conference on Intell. Transportation Systems (ITSC). IEEE, Anchorage, AK, USA.

Potamianos, G., Neti, C., 2000. Stream confidence estimation for audio-visual speech recognition. In: Proceedings of INTERSPEECH. ISCA, Bejing, China.

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

Qin, G., Hotilovac, L., 2008. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. Stat. Methods Med. Res. 17 (2), 207–221.

Rosenberg, A., 2009. Automatic Detection and Classification of Prosodic Events. Columbia University. (Ph.D. thesis).

Rosenberg, A., Hirschberg, J., 2009. Detecting pitch accents at the word, syllable and vowel level. In: Proceedings of The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies: Companion Volume: Short Papers. ACL, Boulder, CO, USA.

Scarborough, R., Keating, P., Mattys, S.L., Cho, T., Alwan, A., 2009. Optical phonetics and visual perception of lexical and phrasal stress in english. Lang. Speech 52 (2−3), 135–175.

Schillingmann, L., Wagner, P., Munier, C., Wrede, B., Rohlfing, K., 2011. Using prominence detection to generate acoustic feedback in tutoring scenarios. In: Proceedings of INTERSPEECH. ISCA, Florence, Italy.

Schnall, A., Heckmann, M., Feature space SVM adaptation for speaker adapted word prominence detection. Submitted to In: Computer Speech and Language.

Schnall, A., Heckmann, M., 2014. Integrating sequence information in the audio-visual detection of word prominence in a human-machine interaction scenario. In: Proceedings of INTERSPEECH. ISCA, Singapore.

Schnall, A., Heckmann, M., 2016. Comparing speaker independent and speaker adapted classification for word prominence detection. In: Proceedings of IEEE Workshop on Spoken Language Technology (SLT). IEEE, San Diego, CA, USA.

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

Schuller, B., Seppi, D., Batliner, A., Maier, A., Steidl, S., 2007. Towards more reality in the recognition of emotional speech. In: Proceedings of IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP). IEEE, Honolulu, HI, USA.

Schwartz, J.-L., Savariaux, C., 2014. No, there is no 150 MS lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. PLoS Comput. Biol. 10 (7), E1003743.

Shriberg, E., 2005. Spontaneous speech: how people really talk and why engineers should care. In: Proceedings of INTERSPEECH. ISCA, Lisbon, Portugal.

Shriberg, E., Stolcke, A., 2004. Direct modeling of prosody: an overview of applications in automatic speech processing. In: Proceedings of Speech Prosody. ISCA, Nara, Japan.

Stiefelhagen, R., Yang, J., Waibel, A., 1997. Tracking eyes and monitoring eye gaze. In: Proceedings of Workshop Perceptual User Interfaces. Banff, Canada.

Streefkerk, B.M., 2002. Prominence. Acoustic and Lexical/Syntactic Correlates. University of Amsterdam (Ph.D. thesis).

Sumby, W., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26 (2), 212–215.

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

Swerts, M., Litman, D., Hirschberg, J., 2000. Corrections in spoken dialogue systems. In: Proceedings of International Conference on Spoken Language Processing (ICSLP). ISCA, Bejing, China.

Tamburini, F., 2003. Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system. In: Proceedings of INTERSPEECH. ISCA, Geneva, Switzerland.

Teissier, P., Robert-Ribes, J., Schwartz, J.-L., Guérin-Dugué, A., 1999. Comparing models for audiovisual fusion in a noisy-vowel recognition task. IEEE Trans. Speech Audio Process. 7 (6), 629–642.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M., 2013. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of 3rd ACM International Workshop on Audio/visual emotion challenge. ACM, Barcelona, Spain.

Varga, A., Steeneken, H., 1993. 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 Commun. 12 (3), 247–251.

Wang, D., Narayanan, S., 2007. An acoustic measure for word prominence in spontaneous speech. IEEE/ACM Trans. Audio Speech Lang. Process 15 (2), 690–701.

Xu, Y., Xu, C.X., 2005. Phonetic realization of focus in English declarative intonation. J. Phonet. 33 (2), 159–197.

Yoshida, T., Nakadai, K., Okuno, H., 2009. Automatic speech recognition improved by two-layered audio-visual integration for robot audition. In: Proceedings of 9th RAS International Conference on Humanoid Robots. IEEE, Paris, France.

You, M., Chen, C., Bu, J., Liu, J., Tao, J., 2006. Emotion recognition from noisy speech. In: Proceedings of IEEE International Conference on Multimedia and Expo. IEEE, Toronto, ON, Canada.

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

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. 31 (1), 39–58.

Zhou, Z., Zhao, G., Hong, X., Pietikäinen, M., 2014. A review of recent advances in visual speech decoding. Image Vis. Comput. 32 (9), 590–605.