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

Environmentally Robust Audio-Visual Speaker Identification

Lea Schönherr, Dennis Orth, Martin Ernst Heckmann, Dorothea Kolossa

2016

Preprint:

This is an accepted article published in IEEE Workshop on Spoken Language Technology. The final authenticated version is available online at: https://doi.org/10.1109/SLT.2016.7846282

© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

DOI: https://doi.org/10.1109/SLT.2016.7846282

ENVIRONMENTALLY ROBUST AUDIO-VISUAL SPEAKER IDENTIFICATION

Lea Schönherr, Dennis Orth, Martin Heckmann, and Dorothea Kolossa

Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany
lea.schoenherr@rub.de, dennis.orth@rub.de, dorothea.kolossa@rub.de
Honda Research Institute Europe GmbH, Germany
martin.heckmann@honda-ri.de

ABSTRACT

To improve the accuracy of audio-visual speaker identification, we propose a new approach, which achieves an optimal combination of the different modalities on the score level. We use the i-vector method for the acoustic and the local binary pattern (LBP) for the visual speaker recognition. Regarding the input data of both modalities, multiple confidence measures are utilized to calculate an optimal weight for the fusion. Thus, oracle weights are chosen in such way as to maximize the difference between the score of the genuine speaker and the person with the best competing score. Based on these oracle weights a mapping function for weight estimation is learned. To test the approach, various combinations of noise levels for the acoustic and visual data are considered. We show that the weighted multimodal identification is far less influenced by the presence of noise or distortions in acoustic or visual observations in comparison to an unweighted combination.

***Index Terms*—** speaker recognition, face recognition, multimodal biometrics, classifier combination, discriminative classifier fusion

1. INTRODUCTION

The performance of speaker recognition, considering acoustic data only, typically shows severe performance impairments in noisy environments. Additionally, audio-only speaker recognition is vulnerable against spoofing attacks, where the attacker claims the identity of another person [1]. Audio-visual speaker recognition can be used to overcome these limitations by extending acoustic speaker recognition with a visual biometric method, e. g., face recognition.

Both modalities can be affected by different kinds of noise: During the visual recognition process, bad lighting conditions, blurring, and rotations of the head can decrease the probability of a successful recognition. In contrast, an acoustic speaker recognition system is not influenced by these specific problems. However, the latter can be compromised by background noise. Therefore, it would be advantageous to combine both systems in such a way that they can compensate each other's weaknesses.

In [2] it has been shown that a combined audio-visual speaker identification can increase the accuracy of the single-modality methods. However, in this approach, only the resulting ranks of the single systems are considered. In other audio-visual speaker recognition systems, the fusion is applied earlier in the process: In [3] an approach for feature fusion is proposed and [4] uses deep Boltzmann machines in combination with deep neural networks to fuse both modalities.

In general, the fusion of multiple classifiers can be characterized by the stage, where the fusion is applied. An early-stage fusion is applied at the feature level. Alternatively, the fusion can be conducted at the decision level, where the features are calculated separately and combined for the final decision. A late-stage fusion, which we use here, is on the score level. Therefore, for each modality a score is calculated and a final decision is obtained by considering the combined scores.

In [5] an overview of fusion strategies for combining multiple modalities is outlined. In our work we will consider the basic product rule. The rule has been considered by other works [6], and it will serve as our baseline.

The main challenge is to combine two sub-systems in such a way that they can truly benefit from each other. Previous work on audio-visual speech recognition [7], [8] has shown that noise-adaptive stream weights can significantly improve the performance of multimodal speech recognition. The idea is to create a side-channel for each modality, which outputs a measure of confidence. This helps to decide how much each modality should contribute to the overall decision. This approach will be described in Section 3 in more detail.

In contrast to [7], in our work, we use the stream weight estimation for an identification task. Thus, we propose a new approach using a new, discriminative cost function to calculate target stream weights, which increase the score of the genuine speaker relative to the most competitive speaker. Further, we introduce confidence measures, which can be used together with the target stream weights in order to learn a mapping function for an optimal weight estimation.

Currently, the basis of many speaker identification meth-

ods are GMM-based speaker models [9], which are also a core component of the i-vector model. The i-vector approach is designed to decouple channel-related and speaker-based signal variabilities and thus offers greater robustness [10].

For a state-of-the-art face recognition we chose the local binary pattern (LBP), due to its ability to distinguish faces effectively [11]. Additionally, the LBP can be implemented with low computational costs and shows high accuracy for changing gray levels [12].

2. SPEAKER AND FACE RECOGNITION

In order to train and test the combination of the acoustic and visual recognition on the score level, the scores for both modalities need to be calculated separately.

2.1. Face Recognition

For face recognition, we chose the LBP, which is essentially based on comparing the value of a pixel with the values of the surrounding pixels. These adjacent pixels are translated into a binary pattern. If the pixel value is smaller, the corresponding position in the binary pattern is set to 0 if the value is larger, the position is set to 1. The resulting binary pattern can be interpreted as an integer value, which is saved for each pixel and is used for the classification. To optimize the recognition, a radius can be defined, describing which neighbors should be considered for each pixel. Additionally, to decrease the dimension, the image is divided into cells and, for each cell, one element for the feature vector $\mathbf{x}_V$ is computed.

Furthermore, the depth images are considered as well. This can later act as a countermeasure against spoofing attacks, using a captured image of the victim. Additionally, and importantly for our application, the depth image is illumination independent, which leads to a more robust recognition. Therefore, both feature vectors computed with LBP are concatenated.

After a training phase, the resulting feature vector of a new image can be used for comparison with all enrolled speakers. At this point, the Euclidean distance is computed between the test image and the mean training image of all speakers C_k with $k = 1, ..., N_S$ to obtain a confusion matrix, where N_S is the number of enrolled speakers .

In addition, for the later audio-visual fusion, a Rayleigh probability density function is fitted onto the distances of the *true positives* to obtain class posterior probabilities $\mathrm{P}(C_k|\mathbf{x}_V)$.

2.2. Speaker Recognition

The i-vector recognition method is more complex than the LBP and involves several steps. We used an implementation for MATLAB by Microsoft Research [13]. In the following we give a brief outline of this approach:

Step 1: In the training phase, MFCCs (mel frequency cepstral coefficients) are extracted from audio files, which contain sentences of the enrolled speakers. This data is used to fit a Gaussian-Mixture-Model-based Universal Background Model (GMM-UBM).

Step 2: The zeroth and first order sufficient statistics (Baum-Welch statistics) for observations (training and test vectors) are computed, given the UBM.

Step 3: Training vectors are created by concatenating the sufficient statistics. These vectors are used to learn the *total variability subspace* (representation of all enrolled speakers) using a factor analysis.

Step 4: Here, the i-vectors for the training and test vectors (in the form of sufficient statistics), are computed with the UBM and the previously calculated *total variability subspace*.

Step 5: The trained i-vectors of all classes are processed using a linear discriminant analysis (LDA) with Fisher's criterion for further dimensionality reduction and to improve the classification in the scoring function.

Step 6: A Gaussian probabilistic LDA is applied to the previously computed training i-vectors. This corresponds to learning a factor analysis model of the i-vectors, which will be used in the actual speaker identification stage.

Step 7: For the actual identification, a log-likelihood ratio is computed, acting as feature vector $\mathbf{x}_A$, and used for a pairwise scoring of the test i-vectors against all enrolled i-vectors. This is done for every possible combination so that a confusion matrix is obtained as the result considering all speakers C_k.

In addition to this confusion matrix, posterior probabilities are needed for the later fusion stages. Thus, an appropriate distribution needs to be fitted onto the likelihood ratios of the *true positives* to obtain $\mathrm{P}(C_k|\mathbf{x}_A)$. Here, Gaussian distributions provided a good fit, and were thus learned on the training data.

3. CLASSIFIER COMBINATION

Once the modality dependent feature vectors $\mathbf{x}_A$ and $\mathbf{x}_V$ have been extracted, both identification systems can compute their respective scores. However, an unweighted combination of the previously introduced scores $\mathrm{P}(C_k|\mathbf{x}_A)$ and $\mathrm{P}(C_k|\mathbf{x}_V)$ is not ideal, because under certain conditions, one of the two systems might be presented with reliable data, whereas the other might only have distorted features available, e. g., due to acoustic noise or low-quality video data.

Therefore, in this work, we suggest to use confidence information, which informs a fusion stage about the reliability of each of the two sub-systems. Thus, the confidence information is utilized to reach a more environmentally robust classification based on noise-dependent weighting of the two subsystems. This approach is explained in further detail in Section 3.2.

3.1. Baseline

As mentioned above, the intention is to compute the probabilities of seeing any of the possible classes, respectively speakers C_k, given the two feature vectors $\mathbf{x}_i$ from the two modalities $i \in \{V, A\}$, where $k = 1 \dots N_S$. Since the probability density $\mathrm{p}(\mathbf{x}_i)$ is unknown, but the likelihood $\mathrm{p}(\mathbf{x}_i|C_j)$ has been learned, we can marginalize over all classes in the denominator:

$$\mathrm{P}(C_k|\mathbf{x}_i) = \frac{\mathrm{p}(\mathbf{x}_i|C_k)\mathrm{P}(C_k)}{\mathrm{p}(\mathbf{x}_i)} = \frac{\mathrm{p}(\mathbf{x}_i|C_k)\mathrm{P}(C_k)}{\sum_{j=1}^{N_S} \mathrm{p}(\mathbf{x}_i|C_j)\mathrm{P}(C_j)}.$$

For each unimodal classifier, the class $C_{\hat{K}}$ is assigned to the input feature vector $\mathbf{x}_i$ if

$$\hat{K} = \arg \max_{k=1..N_S} \mathrm{P}(C_k|\mathbf{x}_i).$$

Applying the same decision rule to the audio-visual fusion task leads to

$$\hat{K} = \arg \max_{k=1..N_S} \mathrm{P}(C_k|\mathbf{x}_V, \mathbf{x}_A). \quad (1)$$

Since the probability $\mathrm{P}(C_k|\mathbf{x}_V, \mathbf{x}_A)$ in Equation (1) is not known, Bayes' theorem and marginalization are applied to rewrite the conditional joint probability, leading to

$$\mathrm{P}(C_k|\mathbf{x}_V, \mathbf{x}_A) = \frac{\mathrm{p}(\mathbf{x}_V, \mathbf{x}_A|C_k)\mathrm{P}(C_k)}{\sum_{j=1}^{N_S} \mathrm{p}(\mathbf{x}_V, \mathbf{x}_A|C_j)\mathrm{P}(C_j)}. \quad (2)$$

Assuming that the feature vectors $\mathbf{x}_V, \mathbf{x}_A$ are statistically independent given the class C_k, the joint distribution can be factorized into

$$\mathrm{p}(\mathbf{x}_V, \mathbf{x}_A|C_k) = \mathrm{p}(\mathbf{x}_V|C_k)\mathrm{p}(\mathbf{x}_A|C_k). \quad (3)$$

Substituting from Equation (3) into Equation (2) and cancelling the priors by considering all speakers as equally likely, the decision rule can be rewritten as

$$\hat{K} = \arg \max_{k=1\dots N_S} \mathrm{P}(C_k|\mathbf{x}_V)\mathrm{P}(C_k|\mathbf{x}_A). \quad (4)$$

3.2. Weighting of Classifiers

A stream weight λ is defined and incorporated in the following decision rule:

$$\hat{K} = \arg \max_{k=1\dots N_S} \mathrm{P}(C_k|\mathbf{x}_V)^{(1-\lambda)}\mathrm{P}(C_k|\mathbf{x}_A)^{\lambda}, \quad (5)$$

such that $0 \leq \lambda \leq 1$. If $\lambda = 0.5$, the decision rule is equivalent to that of unweighted classification. This type of stream weighting has previously led to great accuracy improvements for audio-visual speech recognition, e. g., in [7].

To achieve optimal stream weights for audio-visual identification, we will propose a cost function in Section 3.2.1. This cost function can provide optimal stream weights based on knowing the true speaker identity. Hence, the resulting stream weights can be used as training targets for learning a mapping function f, which uses confidence measures as its input and outputs estimated stream weights.

Thus, we also need appropriate confidence measures. For this purpose, we have considered a range of metrics. Among those, the dispersion D and different estimators of the distortion or noise level of the video and audio files have shown to provide reliable confidence measures. The dispersion is computed over the posterior probabilities obtained for one test file:

$$D_i = \frac{2}{K(K-1)} \sum_{l=1}^{K-1} \sum_{m=l+1}^{K} \log \frac{\mathrm{p}(C_l^*|\mathbf{x}_i)}{\mathrm{p}(C_m^*|\mathbf{x}_i)}, \quad (6)$$

where the K classes $C_1^*, ..., C_K^*$ with the largest probabilities are used, sorted in descending order of likelihood. The value of K can be lower than or equal to the number of enrolled speakers.

The noise level of the audio signals, denoted by ϵ_A, is estimated by a minimum mean-square error log-spectral amplitude estimator [14], for which we have used the MATLAB implementation provided by [15].

To estimate the image distortion (denoted by $\boldsymbol{\epsilon}_V$) three different values are considered, i. e., for each image the lighting condition, the degree of blurring, and the rotation are estimated and used as confidence measures in a vector

$$\boldsymbol{\epsilon}_V = [\epsilon_{V,L}, \epsilon_{V,B}, \epsilon_{V,R}].$$

As the feature $\epsilon_{V,L}$ for the lighting conditions (providing information of whether an image is overexposed or underexposed), the mean pixel value over all pixels is calculated.

A potential blurring, e. g., due to the speaker's movements during image capture, is estimated by applying a Laplacian filter kernel for edge detection. To obtain one feature value $\epsilon_{V,B}$, we calculate

$$\epsilon_{V,Blur} = \sigma^2(I_L),$$

where σ^2 represents the variance and I_L is the image after edge detection.

The last confidence measure $\epsilon_{V,R}$, representing a potential rotation of the speaker's head, is obtained by horizontally mirroring the image and calculating the cross-correlation between the original and the mirrored image. Moreover, to obtain light-independent results for the blurring and rotation measures, $\epsilon_{V,B}$ and $\epsilon_{V,R}$ are computed from the depth image.

Furthermore, we suggest using a function f, which maps all confidence measures to an optimal weight in the sense of our decision rule in Equation (5):

$$\hat{\lambda} = f(D_A, D_V, \epsilon_A, \boldsymbol{\epsilon}_V). \quad (7)$$

Before the mapping function f can be used as given in Equation (7), it has to be learned. For this purpose, we are utilizing supervised machine learning approaches.

To obtain a large number of training targets for learning the mapping function f, both speaker identification systems—audio and video—first need to be trained and in the second step predictions with various data sets have to be calculated. This development set contains N_{DS} files under each of the different acoustic and visual conditions. In our case, N_C = 10 conditions are utilized for each single-modality recognition system. For these data sets, the corresponding dispersion and noise levels are computed. After that, $N = N_{DS} \cdot N_C^2$ input cases for the function f can be formed. In order to learn the mapping function f with these N tuples, we will need training targets, i. e., optimal stream weights for the entire range of the development set.

3.2.1. Optimal Stream Weights

In the following, we suggest an approach to find the optimal stream weights λ_θ for all cases in the development set. The approach leads to ideally discriminative stream weights insofar as it maximizes the ratio of the likelihoods of the true speaker C_{true} and the most likely competing speaker C_{conf}.

For this, we assume C_{true} is the true class of the input feature vector $\mathbf{x}_i$, while C_{conf} is the class, which is most likely to be confused with the true identity, i. e.,

$$C_{\mathsf{conf}} = \arg \max_{\forall C_k \setminus C_{\mathsf{true}}} \mathrm{P}(C_k|\mathbf{x}). \tag{8}$$

Therefore, to find the optimal value of λ_θ, we suggest to maximize the following discriminative cost function for every file in the development set:

$$\begin{aligned} \lambda_\theta &= \arg\max_{\lambda} \left[\frac{\mathrm{P}(C_{\mathsf{true}}|\mathbf{x}_V, \mathbf{x}_A)}{\mathrm{P}(C_{\mathsf{conf}}|\mathbf{x}_V, \mathbf{x}_A)} \right] \mathrm{p}(\lambda) \\ &= \arg\max_{\lambda} [\, \log \mathrm{P}(C_{\mathsf{true}}|\mathbf{x}_V, \mathbf{x}_A) \\ &\quad - \log \mathrm{P}(C_{\mathsf{conf}}|\mathbf{x}_V, \mathbf{x}_A) \,] \, \mathrm{p}(\lambda), \end{aligned} \tag{9}$$

such that $0 \leq \lambda \leq 1$, where the joint distribution probabilities are defined according to (5):

$$\mathrm{P}(C_k|\mathbf{x}_V, \mathbf{x}_A) = \mathrm{P}(C_k|\mathbf{x}_V)^{(1-\lambda)} \mathrm{P}(C_k|\mathbf{x}_A)^{\lambda}. \tag{10}$$

In this way, the distance between the posterior probability of the true and the second class is maximized. If one of the two recognition systems makes a wrong prediction, $\mathrm{P}(C_{\mathsf{conf}}|\mathbf{x})$ will be higher than $\mathrm{P}(C_{\mathsf{true}}|\mathbf{x})$. This effect can typically be mitigated through the choice of better weighting, and an optimal λ is obtained through maximizing (9).

As in [7], we assume that the optimal λ should follow a prior distribution $\mathrm{p}(\lambda) \sim \mathcal{N}(\mu, \sigma^2)$. In our work, μ is obtained during a search, testing different values for μ between 0 and 1 with a step size of 0.01 and using that value, which leads to the highest recognition rate. To refine the result, the variance σ^2 is increased iteratively until all λ_θ become 0 or 1 or a maximum number of iteration steps is reached. Considering all possible λ_θ of all iterations steps, the λ_θ which lead to the best recognition rate are used as *oracle weights*.

Since the true class identity is used for this computation, these stream weights λ_θ are referred to as *oracle weights*. They can thus only serve as training targets for learning f in (7), but are not applicable in practice for speaker identification.

3.2.2. Models for estimating λ

Based on the above considerations, a method for generating training targets λ_θ for the function f is available, but an appropriate model for the function still needs to be chosen. Since the weights λ can be in the range of $[0, 1]$, due to the assumed prior, finding optimum weights becomes a regression problem rather than a classification problem. Experiments have been carried out with *feed-forward neural networks*, either shallow, or deep neural networks (DNN), for the mapping function.

4. EXPERIMENTAL RESULTS

For the experiments, a data set was deployed, which we had recorded with a Kinect sensor from Microsoft. We therefore considered the following data, provided by the Kinect sensor: the four-channel microphone array, the Full HD video, and the captured depth images. The data set contains 30 speakers (15 females and 15 males), with recorded utterances of English digits from 0 to 9. For this work, 4 digits were concatenated randomly for one training or test unit, in order to obtain a longer utterance.

During our experiments, we used all $N_S = 30$ speakers in each phase. This includes the enrollment during the training, the development set to learn the mapping function, and the test set to verify the obtained mapping function. Such a so-called closed-set identification does not consider non-enrolled speakers, like impostors. However, in this work we focus only on the optimal combination of different modalities for speaker identification among enrolled speakers and do not consider impostors.



Fig. 1: For the face recognition different kinds of distortion were considered, i. e., adverse lighting conditions, blurring, and rotations.

In order to train the speaker and face recognition as described in Section 2, a total number of $N_F = 30$ utterances per speaker of the introduced data set were used.

The images were used as gray scale images and processed with the LBP as described in Section 2. The best results for the face recognition could be achieved with a radius of 2 pixels for the video image and a radius of 3 pixels for the depth image. We observed that it is possible to implement a robust face recognition solely using the depth images. Thus, the depth image seems to be a valuable contribution, regarding the robustness and security of face recognition. For our experiments, we used both, the video and the depth images.

For the speaker recognition, the best results were achieved with 24 MFCCs, augmented with their first- and second-order derivatives. The speaker recognition with i-vectors was applied to one channel of the recorded microphone array. To this end, we chose 128 mixture components for the GMM-UBM.

After the training phase, numerous data sets need to be created in order to learn the mapping function. For this purpose, $N_F = 20$ new utterances per speaker were considered for the development set, which had not been used to train the sub-systems. In order to simulate adverse environmental conditions, we added varying amounts of noise to the audio test files and introduced distortions to the video data. For acoustic speaker recognition, white Gaussian noise was added to the utterances. In total, 9 different noisy audio test cases with signal-to-noise ratios (SNRs) between 4 dB and 20 dB were created and the original utterances were also included in the test cases.

For the images, different kinds of distortions were considered, i. e., different lighting, blurring, and rotations. To the depth images, which are independent of the lighting, only the blurring and rotations were applied. For the lighting distortions, we manipulated the pixel values, in order to simulate different conditions. To blur the images, we convolved the image with a filter kernel, which describes the direction and amount of motion for each test case. For this, the kernel is chosen such that the smoothing values are either concentrated in the focus of the kernel (less blurring) or are more distributed in the defined direction (more blurring). In Figure 1 all $N_C = 10$ conditions are shown with the original image located on the bottom right.

Overall, $N_C^2 = 100$ different combinations of acoutic and visual conditions were formed to create the development set. Each condition contains $N_S \cdot N_F = 30 \cdot 20$ recordings. For those combinations, the *oracle weights* and the confidence measure values were computed according to Section 3.2.1.

4.1. Weight Estimation

After the calculation, the values of the *oracle weights* and the confidence measures from the complete development set, are used to train a mapping function. For this purpose, we used *feed-forward neural networks* with different numbers of hidden layers. Additionally, we tested different combinations of confidence measures. For the dispersion calculated by Equation 3.2 we chose $K = 7$.

To assess the performance, the recognition rate R is calulated by

$$R = \frac{N_{\mathsf{rec}}}{N_S \cdot N_F},$$

where N_{rec} denotes the number of correctly classified files. In order to verify the mapping function computed by the NN, a test set, considering $N_F = 20$ new utterances per speaker, was used.

In Table 1 an overview is presented of the different numbers of hidden layers (with 10 neurons per hidden layer) combined with different sets of confidence measures as inputs for the mapping function. Here, the first value in each cell is calculated with the development set (dev) and the second value

Number of Hidden Layers:		1	2	3	4
$f(D_A, D_V, \epsilon_A, \epsilon_V)$	dev	0.939	0.940	**0.941**	0.941
	test	0.932	0.932	**0.932**	0.932
$f(\epsilon_A, \epsilon_V)$	dev	0.936	0.940	0.940	0.940
	test	0.927	0.930	0.929	0.931
$f(D_A, D_V)$	dev	0.928	0.928	0.928	0.928
	test	0.914	0.914	0.914	0.914

Table 1: Recognition rates for different settings.

	λ		4 dB	6 dB	8 dB	10 dB	12 dB	14 dB	16 dB	18 dB	20 dB	clean
			0.30	**0.40**	**0.50**	**0.60**	**0.71**	**0.78**	**0.89**	**0.94**	**0.98**	**1.00**
ID1	0.5	**0.39**	0.35	0.45	0.55	0.66	0.73	0.81	0.91	0.95	0.98	1.00
	$\hat{\lambda}$		0.57	0.68	0.74	0.79	0.83	0.88	0.92	0.96	0.98	1.00
ID2	0.5	**0.45**	0.35	0.43	0.54	0.64	0.74	0.83	0.92	0.96	0.98	1.00
	$\hat{\lambda}$		0.66	0.71	0.77	0.82	0.89	0.93	0.96	0.98	0.99	0.99
ID3	0.5	**0.56**	0.36	0.44	0.54	0.66	0.75	0.84	0.92	0.96	0.98	1.00
	$\hat{\lambda}$		0.74	0.78	0.84	0.89	0.93	0.98	0.99	0.99	0.99	0.99
ID4	0.5	**0.64**	0.35	0.45	0.54	0.65	0.75	0.84	0.92	0.96	0.98	1.00
	$\hat{\lambda}$		0.79	0.83	0.87	0.91	0.94	0.96	0.98	0.98	0.99	0.99
ID5	0.5	**0.68**	0.36	0.45	0.55	0.66	0.74	0.82	0.91	0.95	0.98	1.00
	$\hat{\lambda}$		0.74	0.79	0.83	0.86	0.90	0.94	0.97	0.98	0.99	1.00
ID6	0.5	**0.76**	0.35	0.45	0.55	0.66	0.74	0.84	0.92	0.96	0.98	1.00
	$\hat{\lambda}$		0.85	0.89	0.93	0.96	0.98	0.98	0.99	0.99	1.00	1.00
ID7	0.5	**0.85**	0.36	0.46	0.56	0.67	0.74	0.83	0.91	0.96	0.98	1.00
	$\hat{\lambda}$		0.88	0.90	0.93	0.96	0.98	0.99	1.00	1.00	1.00	1.00
ID8	0.5	**0.87**	0.36	0.46	0.55	0.66	0.75	0.84	0.92	0.96	0.98	1.00
	$\hat{\lambda}$		0.92	0.96	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00
ID9	0.5	**0.98**	0.36	0.44	0.54	0.66	0.74	0.83	0.92	0.96	0.98	1.00
	$\hat{\lambda}$		0.97	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
OI	0.5	**1.00**	0.35	0.44	0.53	0.64	0.73	0.81	0.91	0.95	0.98	1.00
	$\hat{\lambda}$		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2: Recognition rates for every combination of audio noise and video distortions achieved on the test set. The rows present the different image distortions (ID1 – ID9) and the original image (OI). The different audio test cases are presented in the columns. In the gray cells, the recognition rates for the single-modality systems are shown.

with the test set. For this, we always used the output $\hat{\lambda}$ of the mapping function, trained using the confidence measures as denoted in the first column.

By comparing the results of the development set and the test set, one can see that the mapping function remains robust for unseen data. Further, using only the dispersion values or the estimated noise/distortion levels leads to almost equally high recognition rates as using all confidence measures. Moreover, changing the number of layers does not substantially affect the recognition rate. However, using 3 hidden layers and all confidence measures led to the highest accuracy for combinations using clean audio and video data and was therefore chosen for the following experiments.

In Table 2 the results for this setting (bold values in Table 1) are shown in detail. Here, all $N_C^2 = 10 \cdot 10$ possible combinations of the test conditions are presented. In the gray cells, the recognition rate of the single-modality systems are displayed. In the remaining cells, the combined recognition is shown. Here, the top value represents the recognition rate obtained with the baseline approach ($\lambda = 0.5$). The second value is the recognition rate achieved with the weights $\hat{\lambda}$, estimated by the DNN.

As one can see, in the results for the baseline system with $\lambda = 0.5$, the audio recognition shows high recognition rates for test cases with high SNR and low recognition rates for test cases with low SNR. On the other hand, the combined recognition rate barely seems influenced by the face recognition. This observation is consistent with the consideration, that in an unweighted recognition, the result depends on the distribution of the scores of the single systems.

When the noise level increases one would expect a classifier to yield to a flat posteriori probability distribution, as it will not be able to reliably differentiate between the classes. Yet, classifiers applied outside of their training domain, i. e., trained on clean data and applied on noisy data, have a tendency to yield peaked distributions, preferring certain classes [8]. This contradicts the assumption of the purely Bayesian fusion for $\lambda = 0.5$ and yields inferior results for the unweighted fusion with increasing noise levels (compare Table 2). The weighting of the streams counteracts this effect by introducing an external signal for classifier confidence.

Therefore, if both modalities are weighted equally, a correct classification in the face recognition is not able to compensate a wrong classification in the speaker recognition.

In contrast, the mapping function, obtained by the DNN, performs just as expected: It increases all recognition rates up to at least very close to the highest recognition rate of the single-modality systems. For a few combinations at high SNRs, there are slight decreases in comparison to the baseline result. However, the recognition rate in these cases remains very high and the changes in accuracy never exceed 1 %.

On the whole, for the complete test set, the average improvement was from 74.17 % to 93.23 %, which shows the applicability of the presented approach also in those situations where stream weights are not based on oracle information, but rather estimated from the newly suggested approach, based on easily estimated confidence values.

5. CONCLUSIONS

We have shown that a better speaker identification can be obtained by a fusion of state-of-the-art audio-based and video-based identification. For this purpose, we have proposed a new weighting approach for the two modalities using a discriminative cost function to increase the ratio of the score of the true speaker relative to the speaker with the most competitive score. *Feed-forward neural networks* with multiple hidden layers led to the best results for computing these stream weights, based on a set of confidence measures.

We have observed that with a multimodal recognition it is fairly easy to deal with different kinds of noises or distortions added to the audio and video data by using an estimation of these distortions together with the dispersion of the scores.

On the whole, a large improvement over all considered test cases can be observed. Additionally, the results are independent of the posteriori probability distribution of the single-modality recognition systems. Importantly, in all considered cases, we achieved at least principally the recognition rate we would achieve with the best single modality. Moreover, for the test cases with low recognition rates, we were always able to exceed these values, in many cases very notably.

6. REFERENCES

[1] G. Chollet, P. Perrot, and W. Karam, "Identities, forgeries and disguises," *Int. J. Inform. Manage.*, 2012.

[2] M. Alam, R. Togneri, and F. Sohel, "Linear Regression-based Classifier for Audio Visual Person Identification," *ICCSPA*, 2013.

[3] C. Yu and L. Huang, "Biometric Recognition by Using Audio and Visual Feature Fusion," *ICSSE*, 2012.

[4] M. Alam, M. Bennamoun, and R. Togneri, "A Deep Neural Network for Audio-Visual Person Recognition," *BTAS*, 2015.

[5] P. Atrey, M. Hossain, and A. El Saddik, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, 2010.

[6] J. Kittler, M. Hatef, and R. Duin, "On combining classifiers," *IEEE T. Pattern Anal. Mach. Intell.*, 1998.

[7] A.H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning Dynamic Stream Weights For Coupled-HMM-Based Audio-Visual Speech Recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2015.

[8] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise Adaptive Stream Weighting in Audio-visual Speech Recognition," *EURASIP J. Appl. Signal Process.*, 2002.

[9] D. A. Reynolds and R. C. Rose, "Robust Text-independent Speaker Identification using Gaussian Mixture Speaker Models," *IEEE T. Speech Audi. P.*, 1995.

[10] N. Dehak, P. J. Kenny, and R. Dehak, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. Audio, Speech, Language Process.*, 2011.

[11] D. Huang, C. Shan, and M. Ardabilian, "Local Binary Patterns and Its Application to Facial Image Analysis: A Survey," *IEEE Trans. Syst., Man, Cybern., Syst*, 2011.

[12] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Trans. Pattern Anal. Mach. Intell*, 2006.

[13] S.Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker-Recognition Research," *Speech and Language Processing Technical Committee Newsletter*, 2013.

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE T. Acoust. Speech*, 1985.

[15] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2013.