A COMPACT FORMULATION OF TURBO AUDIO-VISUAL SPEECH RECOGNITION

Simon Receveur, Patrick Meyer, and Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig
38106 Braunschweig, Germany
Email: {s.receveur, patrick.meyer, t.fingscheidt}@tu-bs.de

ABSTRACT

Since most automatic speech recognition (ASR) systems still suffer from adverse acoustic conditions and insufficient acoustic modeling, recognition robustness can be improved by integrating further information sources such as additional acoustic channels, modalities, or models. Considering the question of information fusion, interesting parallels to problems in digital communications can be observed, where the turbo principle revolutionized reliable communication. In this paper, we provide new perspectives on *turbo ASR*: First, we introduce a *compact* formulation of turbo automatic speech recognition; second, we present a shape-based visual feature extraction algorithm without any learning paradigms. Third, we show an application to an audio-visual speech recognition task on a large data set, where our proposed method clearly outperforms the iterative approach introduced by Shivappa et al. as well as a conventional coupled-hidden-Markov-model approach by up to 23.8% relative reduction in word error rate.

Index Terms— Multimedia systems, speech recognition, iterative decoding, hidden Markov models

1. INTRODUCTION AND PRIOR WORK

In 1993 the so-called *turbo codes* invented by Berrou et al. [1] innovated communication theory. Based on very simple component coding schemes (parallel or serial), they showed how to approach the theoretical performance bounds. One of the virtues of the turbo principle lies in the ability of a highly efficient decoding, applying an iterative processing with simple component decoders. In this decoding process, local reliability estimates are provided, which are utilized in an iterative information fusion. The decoding algorithm providing such *soft information* in the form of state posterior probabilities is, e.g., the BCJR algorithm [2]. In automatic speech recognition, the BCJR algorithm is better known as *forward-backward-algorithm* (FBA) and can be used for recognition. Considering these parallels, the question arises whether the immense gains obtained in communications could also probably be achieved in the field of ASR. This subject forms the focus of the work presented here.

Despite the commercial success and widespread application, most ASR systems still perform poorly in adverse acoustic conditions e. g., background noise or channel distortions. However, their robustness can be improved by exploiting further information sources such as additional acoustic channels [3, 4], modalities [5–7], or models [8, 9]. Here, the success of such approaches is closely linked to the used method of information fusion; considering hidden Markov model (HMM) classifiers, commonly data-dependent combination functions such as the weighted product rule are used for this purpose [10, 11]. In particular, the joint probability distribution of the observation likelihoods is normally composed by means of a weighted product of the individual observation likelihoods extracted by the respective HMM classifiers. Thereby, the relative influence of each information source or *stream* is controlled by a weighting parameter or so-called *stream weight*, e. g., according to its reliability [12]. In the field of ASR, such stream weights were first applied to speech-noise decomposition [13] and later on were adopted to multi-band audio-only ASR [3] as well as to audio-visual ASR [7]. Current decision fusion approaches for audio-visual ASR such as coupled [14, 15] or multi-stream HMMs [12] still incorporate such a weighting scheme while computing joint observation likelihood distributions; gauged on their ASR performance, the determination of appropriate weights is of central importance. In contrast, when applying conventional late fusion techniques such as confusion networks or ROVER [16] in multimodal ASR, the information fusion benefits are limited to the locally best output segment in the combined word transition network. Moreover, reliable word confidence estimates required in the subsequent voting schemes are often difficult to obtain.

Considering the parallels of communications and ASR, Shivappa et al. introduced an iterative approach to multimodal ASR [17], which solves the fusion problem by an iterative recognition scheme. Interestingly, their approach does not employ any stream weighting, but still has the advantage of separately trained HMMs for each modality instead of a joint one. However, during recognition the iterative decoding is controlled by a rate parameter while modeling and re-estimating the distributions of the observation likelihoods [18]. Originating from Shivappa's altered FBA approach, we showed in our previous work [19] that the unmodified FBA is already suitable for iterative recognition. This can be achieved by modifying the observation likelihoods to allow injection of information from a previous iteration. Moreover, the solution required no modeling of observation likelihood distributions. Accordingly, we extended our previous work to a generalized *turbo ASR* approach, which is fully applicable to single- and multi-channel ASR, single- and multi-modal ASR, as well as to single- and multi-model ASR.

In this paper, we introduce the generalized turbo ASR approach in a compact vector-matrix notation, allowing a clear view on differences to Shivappa's approach [17, 18]. Applied to audio-visual speech recognition on a large data set, we use a novel shape-based visual feature extraction algorithm, which dispenses with the commonly used learning paradigms.

The organization of the paper is as follows: In Section 2, we introduce the turbo recognition approach in a compact formulation. Section 3 presents the advantageous audio-visual feature extraction. In Section 4, we report on the performance of turbo ASR on an audio-visual speech recognition task compared with iterative and conventional information fusion methods. The paper concludes with Section 5.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

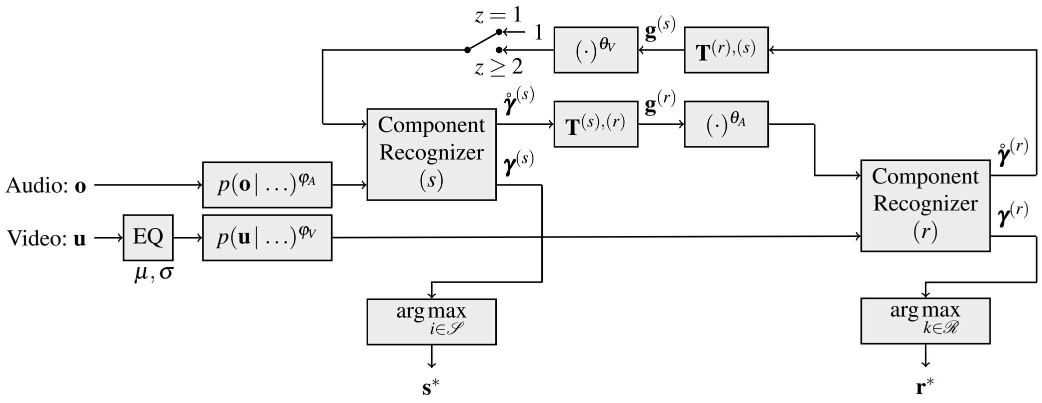
$z=1$
1
$z\geq 2$
$(\cdot)^{\theta_V}$
$\mathbf{g}^{(s)}$
$\mathbf{T}^{(r),(s)}$
$\mathring{\boldsymbol{\gamma}}^{(s)}$
$\mathbf{T}^{(s),(r)}$
$\mathbf{g}^{(r)}$
$(\cdot)^{\theta_A}$
Component Recognizer (s)
$\boldsymbol{\gamma}^{(s)}$
Audio: $\mathbf{o}$
$p(\mathbf{o}\,|\,\ldots)^{\varphi_A}$
Video: $\mathbf{u}$
EQ
μ,σ
$p(\mathbf{u}\,|\,\ldots)^{\varphi_V}$
Component Recognizer (r)
$\mathring{\boldsymbol{\gamma}}^{(r)}$
$\boldsymbol{\gamma}^{(r)}$
$\arg\max_{i\in\mathscr{S}}$
$\arg\max_{k\in\mathscr{R}}$
$\mathbf{s}^*$
$\mathbf{r}^*$

Fig. 1. Turbo audio-visual speech recognizer with iteration index $z = 1,2,\ldots$, starting with the audio stream; time index t omitted.

2. THE TURBO ASR APPROACH

2.1. Notations

Let $\mathbf{x}_1^T = \mathbf{x}_1,\ldots,\mathbf{x}_T$ be a sequence of d_o-dimensional feature vectors with values $\mathbf{x}_t = \mathbf{o}_t \in \mathbb{R}^{d_o}$ for each frame $t = 1,\ldots,T$. This feature vector sequence is supplied to a speech recognizer utilizing an HMM $\lambda = \{\boldsymbol{\pi};\mathbf{A};\mathscr{B}\}$, whose parameters are given by $\boldsymbol{\pi} = [\pi_1,\ldots,\pi_N]^\mathsf{T}$, the vector of prior probabilities $\pi_i = \mathrm{P}(s_1{=}i)$ of all states $i \in \mathscr{S} = \{1,\ldots,N\}$, $\mathbf{A} = \{a_{j,i}\}_{j,i\in\mathscr{S}}$, the matrix of state transition probabilities $a_{j,i} = \mathrm{P}(s_t{=}i\,|\,s_{t-1}{=}j)$, and $\mathscr{B} = \{b_i(\mathbf{x}_t)\}_{i\in\mathscr{S}}$, the set of d_o-variate emission probability density functions (pdfs) $b_i(\mathbf{x}_t) = \mathrm{p}(\mathbf{x}_t\,|\,s_t{=}i)$. The latter may also be expressed in vectorial notation as $\mathbf{b}_t = [b_1(\mathbf{x}_t),\ldots,b_N(\mathbf{x}_t)]^\mathsf{T}$, with $[\,]^\mathsf{T}$ being the transpose. Note that we use $\mathrm{P}(\cdot)$ for probabilities and $\mathrm{p}(\cdot)$ for pdfs (or their values).

Now let there be another observation sequence $\mathbf{u}_1^T$ from a discriminative feature space $\mathbb{R}^{d_u}$. Its d_u-dimensional feature vectors shall be of the same length T as $\mathbf{o}_1^T$ and originate from another sensor (same or different modality)[1]. Furthermore let there be two state-level *maximum-a-posteriori* (MAP) recognizers concatenated in parallel as sketched in Figure 1. Each of these component recognizers (CRs) processes one of the given (discriminative) feature sequences $\mathbf{o}_1^T$ and $\mathbf{u}_1^T$ and employs an individually trained HMM matching the incoming observations. For distinction, the two CRs shall be denoted with (s) and (r). Let the CR (s) be linked with the feature sequence $\mathbf{o}_1^T$ employing an HMM $\lambda^{(s)}$, while the feature sequence $\mathbf{u}_1^T$ is processed by the CR (r) incorporating an HMM $\lambda^{(r)}$. Note that we apply the superscripts (s) and (r) labeling the respective state index spaces $\mathscr{S} = \{1,\ldots,N\}$ and $\mathscr{R} = \{1,\ldots,M\}$.

2.2. Turbo Forward-Backward Algorithm (FBA)

Given the feature vector sequence $\mathbf{o}_1^T$, at each time $t = 1,\ldots,T$ each HMM state $i \in \mathscr{S}$ is linked to a posterior probability $\gamma_t(i) = \mathrm{P}(s_t{=}i\,|\,\mathbf{o}_1^T)$. Using an iterative recognition approach [19], the vector of state posteriors $\boldsymbol{\gamma}_t^{(s)} = [\gamma_t^{(s)}(1),\ldots,\gamma_t^{(s)}(N)]^\mathsf{T}$ is obtained by

$$\boldsymbol{\gamma}_t^{(s)} = \frac{1}{C_t}\cdot\left[\boldsymbol{\alpha}_t^{(s)} \circ \boldsymbol{\beta}_t^{(s)}\right], \tag{1}$$

$$\boldsymbol{\alpha}_t^{(s)} = \mathbf{b}_t^{(s)} \circ \mathbf{g}_t^{(s)} \circ \left[\mathbf{A}^{(s)}\cdot\boldsymbol{\alpha}_{t-1}^{(s)}\right], \tag{2}$$

$$\boldsymbol{\beta}_t^{(s)} = \mathbf{A}^{(s)\mathsf{T}}\cdot\left[\mathbf{b}_{t+1}^{(s)} \circ \mathbf{g}_{t+1}^{(s)} \circ \boldsymbol{\beta}_{t+1}^{(s)}\right], \tag{3}$$

where the *forward* and *backward variables* $\alpha_t(i) = \mathrm{p}(\mathbf{o}_1^t, s_t{=}i)$ and $\beta_t(i) = \mathrm{p}(\mathbf{o}_{t+1}^T\,|\,s_t{=}i)$ for all $i \in \mathscr{S}$ are denoted in vectorial notations $\boldsymbol{\alpha}_t^{(s)} = [\alpha_t^{(s)}(1),\ldots,\alpha_t^{(s)}(N)]^\mathsf{T}$ and $\boldsymbol{\beta}_t^{(s)} = [\beta_t^{(s)}(1),\ldots,\beta_t^{(s)}(N)]^\mathsf{T}$. These variables are initialized to $\boldsymbol{\alpha}_1 = \boldsymbol{\pi}\circ\mathbf{b}_{t=1}$ and $\boldsymbol{\beta}_T = \mathbf{1}_{N\times 1}$, with $\mathbf{1}_{N\times 1}$ being an $N\times 1$-dimensional vector containing ones; the $(\circ)$ operator marks the element-wise product. Then computation is done recursively according to (2) and (3). Note, that vector $\mathbf{g}_t^{(s)} = [g_t^{(s)}(1),\ldots,g_t^{(s)}(N)]^\mathsf{T}$ indicates the *extrinsic* information to be passed on between the CRs. Moreover, the stochastic constraint in (1) is ensured by the normalization $C_t = \boldsymbol{\alpha}_t^{(s)\mathsf{T}}\cdot\boldsymbol{\beta}_t^{(s)}$.

Given state-level MAP recognizers, the sequence $\mathbf{s}^* = s_1^*,\ldots,s_T^*$ of (locally) most probable states is provided by

$$s_t^* = \arg\max_{i\in\mathscr{S}} \boldsymbol{\gamma}_t, \qquad t = 1,\ldots,T. \tag{4}$$

Analogous to the turbo principle, in addition to the feature vector sequences soft state information is passed between the CRs (Figure 1). Due to stability criteria [19, Section IV] thereby the M-dimensional vector of fed back *extrinsic probabilities* $\mathring{\boldsymbol{\gamma}}_t^{(r)} = [\mathring{\gamma}_t^{(r)}(1),\ldots,\mathring{\gamma}_t^{(r)}(M)]^\mathsf{T}$ from recognizer (r) is related, but not equal to the vector of state posteriors $\boldsymbol{\gamma}_t^{(r)}$ of CR (r); the other direction is accordingly.

When considering the information fusion within each recognizer, Shivappa et al. regarded the extrinsic probabilities $\mathring{\boldsymbol{\gamma}}_t^{(r)}$ as an additional observation vector independent of $\mathbf{o}_1^T$ to be fed into CR (s) and deduced a modified FBA [17]. However, it can be shown that it is sufficient to modify the emission terms of the respective HMMs only to allow an injection of extrinsic information of a previous iteration [19]. Thus, the fed-back extrinsic information vector is given by

$$\mathring{\boldsymbol{\gamma}}_t^{(s)} = \frac{1}{C_t'}\left[\mathbf{A}^{(s)}\cdot\boldsymbol{\alpha}_{t-1}^{(s)}\right]\circ\boldsymbol{\beta}_t^{(s)}, \tag{5}$$

with C_t' ensuring the stochastic constraint. Please note that in contrast to [17] in (5) the channel or *intrinsic* information of current frame t is removed.

Eq. (5) is based on the assumption of equal HMM state index spaces within each CR. However, the respective state index spaces $\mathscr{R}$ and $\mathscr{S}$ may differ in multisensor ASR systems, e. g., audio-visual speech recognition. We take account of this fact by merely assuming a known prior co-occurrence probability for all HMM states $i \in \mathscr{S}$ and $k \in \mathscr{R}$. Using the *linear transformation* matrix

$$\mathbf{T}^{(r),(s)} = \{T_{k,i}^{(r),(s)}\}_{k\in\mathscr{R},i\in\mathscr{S}} = [\mathbf{T}^{(s),(r)}]^\mathsf{T} \tag{6}$$

[1] In the special case that $\mathbf{u}_1^T = \mathbf{o}_1^T$, the turbo method could still be applied by using two different recognizers or HMMs.

to relay the extrinsic probabilities $\mathring{\boldsymbol{\gamma}}_t^{(r)}$ from state index space $\mathscr{R}$ to $\mathscr{S}$, the extrinsic information $\mathbf{g}_t^{(s)}$ to be passed between the CR (r) and CR (s) is given by

$$\mathbf{g}_t^{(s)} = \mathbf{T}^{(r),(s)} \cdot \mathring{\boldsymbol{\gamma}}_t^{(r)}. \tag{7}$$

Thereby the respective linear state index transformation matrix elements are given by

$$T_{k,i}^{(r),(s)} = \frac{\mathrm{P}(r_t = k, s_t = i)}{\mathrm{P}(r_t = k)\,\mathrm{P}(s_t = i)}, \ \forall (i,k) \in \mathscr{S} \times \mathscr{R}. \tag{8}$$

The joint probabilities needed in (8) can be determined by using a reference FBA to compute the state posteriors $\bar{\gamma}_\tau^{(s)}(i)$ and $\bar{\gamma}_\tau^{(r)}(k)$ on training data and subsequently estimate a joint probability according to

$$\widehat{\mathrm{P}}(r_t = k, s_t = i) = \frac{1}{C''} \sum_{\tau} \bar{\gamma}_\tau^{(s)}(i) \bar{\gamma}_\tau^{(r)}(k), \ \forall (i,k) \in \mathscr{S} \times \mathscr{R}. \tag{9}$$

Here, the normalization C'' ensures the stochastic constraint and the sum is taken over all training frames τ. Moreover, the prior probabilities $\mathrm{P}(r_t = k)$ and $\mathrm{P}(s_t = i)$ are obtained by marginalization.

In order to ensure convergence in information fusion, the emissions $\mathbf{b}_t$ and the extrinsic information $\mathbf{g}_t$ need to fulfill some numerical prerequisites. First of all, the emissions of both streams should have a similar numeric range, which is rarely the case in multimodal ASR. A simple yet effective method is a histogram equalization: During training, the means μ_s, μ_r and standard deviations σ_s, σ_r of the respective emissions $b_i^{(s)}, b_k^{(r)}$ are estimated; prior to recognition, the emissions of one of the streams are equalized to match the histogram of the other, as marked by the "EQ" block in Figure 1:

$$\overline{b}_k^{(r)}(\mathbf{u}_t) = \left(b_k^{(r)}(\mathbf{u}_t) - \mu_r \right) \cdot \frac{\sigma_s}{\sigma_r} + \mu_s. \tag{10}$$

The balance between emissions and extrinsic information can be actively influenced using a weighting scheme as, e. g., in a coupled HMM. By weighting the emissions, a constant bias in the reliability of the respective modality can be adjusted [7, 12], e. g., depending on the signal-to-noise-ratio (SNR) [14, 15]. Accordingly, we incorporate likelihood weights $0 \le \varphi_A, \varphi_V \le 1$ for audio and video emissions, respectively. Moreover, likelihood weights on the extrinsic information are used to adjust its peakedness. The latter are especially important for controlling the convergence behavior: We found that as the number of iterations z increases, gradually shifting the influence from the emissions towards the extrinsic information improves convergence behavior. Thus, we utilize two extrinsic weights θ_A, θ_V that grow dynamically according to a logistic function

$$\theta(z) = \frac{1}{1 + \mathrm{e}^{-\rho(z-2)} \left(\frac{1}{\theta(2)} - 1 \right)}, \qquad z = 2, 3, \ldots, \tag{11}$$

with $\theta(z) \in \{\theta_A(z), \theta_V(z)\}$. Here, $\theta(2) \in \{\theta_A(2), \theta_V(2)\}$ and $\rho \in \{\rho_A, \rho_V\}$ mark the initial extrinsic weight and the logistic proportionality constant, respectively. Hence, beginning from a given initial value, the extrinsic weights yield to unity as the number of iterations z increases. Please refer to Figure 1 for a summary of the entire turbo recognition scheme.

3. AUDIOVISUAL FEATURE EXTRACTION

3.1. Face and Mouth Detection

Commonly, visual feature extraction algorithms for shape-based features require an offline training step using a number of previously labeled frames [7, Sect. 3]. While exploiting large databases with various speakers, the effort of such a mouth-labeling step by hand for detection purpose becomes enormous. To avoid this considerable burden, we developed a generalized signal processing-based algorithm extracting shape-based features, which dispenses with typically utilized learning paradigms: First, color segmentation in the HSV color space is performed to locate skin-like areas. After grouping these areas and applying a chain of morphological operators to enhance the connected regions, the largest area is assumed as the face candidate. Within this region, eye and mouth candidates are estimated individually by using hybrid methods: The eye localization combines color, edge and illumination-based approaches [20], whereas the mouth candidates are determined by using an edge [21] and color-based technique [22]. Based on these hybrid approaches, different weightings can be assigned for each candidate reflecting the number of methods providing evidence. Given the weighted candidates, we hierarchically choose the set that best matches a given model of a face [21].

3.2. Visual Feature Extraction

The goal of our feature extraction is to describe the precise shape of the lips with a small number of coefficients. Thus, contours of the upper and lower lip are required, which are determined by combining edge detection [21] and lip-color transformation [23] techniques on the assumed mouth region. At first, horizontal filtering is applied on a gray scale image highlighting bright-to-dark intensity changes (top down). The resulting edge image is weighted by means of a lip color transformation map. Subsequently, an upper lip contour hypothesis is obtained by an iterative threshold determination to binarize the processed image. Based on this upper lip estimate, the potential position of the lower lip is further narrowed down by applying an edge detection. The shape of the lower lip is subsequently obtained through an iterative threshold calculation to binarize the lip color transformation image in the defined region. Finally, the lips are aligned to each other and the best positions for the lip corners are determined. Once both lips are obtained, 18 points are set along the center line of each lip with the same horizontal distance, whereby the lip corners constitute the outer boundaries. The coordinates of these 36 points are the requested features. Each set of coordinates is aligned through procrustes analysis [24]. This usual alignment of shape-based features requires an average model of a mouth, which is generated by an offline training process in advance. Additionally, to diminish redundancy and dimensionality, we apply a principal component analysis and regard 98 % of the variance, while reducing the number of coefficients from 72 to 11 in this work.

3.3. Acoustic Feature Extraction

The acoustic features are computed according to the ETSI Advanced Front-End (AFE) Recommendation [25] from 8 kHz audio data, applying a Hamming window of length 25 ms and a frame shift of 10 ms. In conclusion, the produced feature vectors consist of 40 coefficients, with 13 MFCC coefficients, 1st- and 2nd-order derivatives, and additionally one log energy parameter.

4. EVALUATION

4.1. Compared ASR Systems

We applied the iterative decoding approach introduced by Shivappa et al. [17] as audio-visual ASR reference. Within that approach, we estimated the variance of the likelihood values during recognition stage at each iteration and used the result as rate parameter $\frac{1}{\rho}$, while iteratively re-estimating the distribution of the observation likelihood values. Moreover, to improve fairness of comparison to our turbo FBA partly using SNR-dependent weights, we even improved the exponential distribution within the iterative approach by introducing an additional SNR-dependent exponential scaling factor υ_{SNR} being optimized separately in advance.

Moreover, a conventional coupled-HMM (CHMM) approach was employed as a decision fusion baseline [7, 14]. The CHMM system utilizes the weighted product rule [10, 11] for fusion, which incorporates two exponential stream weights φ'_A and φ'_V on the audio and video emissions, respectively. The weights are separately optimized during training—φ'_A as SNR-dependent—, letting $0 \leq \varphi'_A, \varphi'_V \leq 1$ and $\varphi'_A + \varphi'_V = 1$.

4.2. Experimental Setup

We apply the presented turbo FBA approach to a speaker-dependent audio-visual ASR task. All experiments are based on the GRID audio-visual speech corpus containing audio and video recordings of 1000 utterances per speaker [26].

We selected 20 (10 male and 10 female) speakers for the experiments reported here, whereas 4 (2 male and 2 female) additional speakers were employed for parameter training. Moreover, the audio recordings were interfered with white Gaussian noise at fixed SNRs (0 dB up to 30 dB active speech level, 5 dB steps) based on *ITU-T P.56* [27]. For each speaker, 800 randomly chosen utterances are used for HMM training; the remaining 200 utterances are used in the test set. We trained speaker-dependent HMMs separately for each CR (video or undisturbed audio). Each HMM set comprised 51 word HMMs (according to the GRID vocabulary) with a linear topology, using a rule of four emitting states per phoneme. The state emission pdfs were modeled with Gaussian mixture models of order 5 and diagonal covariance matrices.

The following parameters were optimized separately on the test data of the 4 parameter training speakers; in the test stage, the found parameters were adopted for the 20 evaluation speakers. Thus, for the SNRs from 0 dB to 30 dB we obtained the CHMM stream weights $\varphi'_A = (0.1, 0.15, 0.9, 0.95, 0.95, 0.95, 0.95)$ and the turbo FBA emission weights $\varphi_A = (0.14, 0.25, 0.7, 0.7, 0.83, 0.83, 0.92)$. As in [7, 14], we constrained the video CHMM stream weight by $\varphi'_V = 1 - \varphi'_A$, while in our turbo system the video emission weight was set to a fixed $\varphi_V = 0.01$. In the first two iterations, however, we set the emission weights φ_A and φ_V to unity ensuring reference FBA behavior. For the extrinsic weights, we obtained the initial values $\theta_A(2) = 10^{-5}$, $\theta_V(2) = 0.2$ and the logistic proportionality constants $\rho_A = 2.5, \rho_V = 0.65$. Moreover, we attained the iterative scaling factors $10^3 \cdot \upsilon_{SNR} = (7.9, 125, 15.6, 31.3, 62.5, 125, 500)$.

For each SNR, we carried out eight turbo iterations and computed the output posteriors of each CR. As a performance measure, we used the word recognition accuracy in percent, given by $\text{ACC} = \frac{N-D-I-S}{N}$, where N, D, S, I mark the number of reference labels, deletions, substitutions, and insertions, respectively. For this measure to be applicable, we converted the MAP state sequences to word sequences by first allocating each state in the sequence to the respective word identity of its containing word HMM and then merging strings of consecutive identical words. This can be easily done due to the surjective relation between a state and the word identity.

4.3. Results

Figure 2 illustrates the results of our recognition experiments. The dotted lines with triangular markers show the single-channel baselines for audio (△) and video (▽), using a reference FBA. Furthermore, the dotted line with (◇) markers plot the audio-visual CHMM baseline. The remaining lines with ($*$) and ($\circ$) markers indicate the recognition results of the iterative reference (dashed lines) and the herein presented turbo FBA (solid lines): the curve with ($*$) markers was obtained by starting with the audio CR in the first iteration and then examining the output of both CRs in an alternating fashion. Analogously, the ($\circ$) marked curve was generated by starting with the video CR.

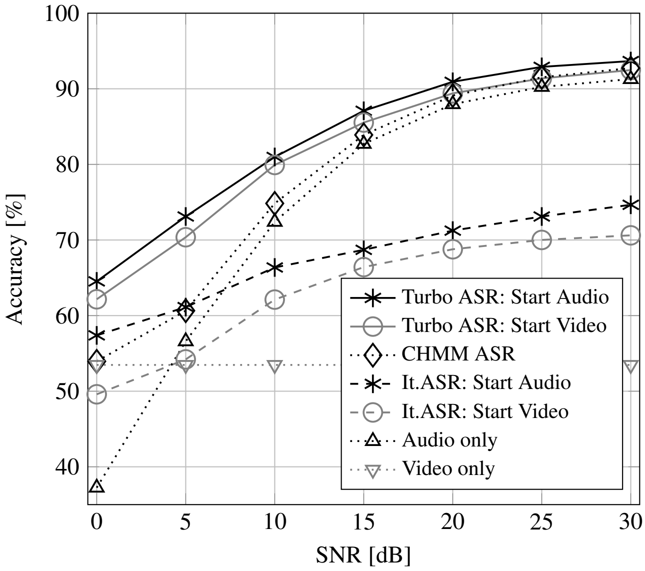

Fig. 2. Recognition results in word accuracy (% ACC) vs. SNR (dB). The dotted lines with triangular markers represent single-channel baselines (△: audio, ▽: video), the dotted line with (◇) markers illustrate a conventional audio-visual CHMM approach [7, 14]. The lines with ($*$) and ($\circ$) markers indicate the recognition results of the iterative reference (dashed lines) [17] and the herein presented turbo FBA (solid lines), both after the eighth iteration, starting with the audio or video CR, respectively.

The following single-modality accuracies were achieved: 53.5% on the video-only test corpus, while the audio-only recognition results vary from 37.2 % at 0 dB SNR to 91.3 % at 30 dB SNR. In comparison, the audio-visual CHMM approach yields recognition results of 53.9 % at 0 dB SNR up to 92.7 % at 30 dB SNR, serving as a sound reference. The iterative reference incorporating a parametric model (reinforcing selectively the most probable state) does not perform convincingly in this context, which might be due to a high dependency on the chosen features, as stated by the authors [18, Sec. 4.2]. The new turbo FBA, however, brought significant improvements, outperforming the iterative baseline by at least 4.8 % for all SNRs. In addition, the quite strong CHMM approach is exceeded over the whole SNR range. At 5 dB SNR, the turbo ASR system outperforms the iterative as well as the CHMM baseline by about 9.2 % absolute, which corresponds to a relative word error rate (WER) reduction of 23.8 %.

5. CONCLUSIONS

In this paper we proposed a compact formulation of a *turbo*-decoding forward-backward algorithm, which is fully applicable to single- and multichannel ASR. Applied to an audio-visual speech recognition task on a large data set, we presented a shape-based visual feature extraction, which dispenses with commonly needed learning paradigms. The experimental results showed that our proposed method clearly outperforms both known iterative and conventional information fusion methods by a relative WER reduction up to 23.8%. For future work, this paper paves the way for principal investigations on further information sources such as additional acoustic channels or models.

6. REFERENCES

[1] Berrou, C.; Glavieux, A.; Thitimajshima, P., "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes," in *Proc. of IEEE Int. Conf. on Communications (ICC 1993)*, Geneva, Switzerland, May 1993, pp. 1064–1070.

[2] Bahl, L.; Cocke, J.; Jelinek, F.; Raviv, J., "Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 284–287, Mar. 1974.

[3] Bourlard, H.; Dupont, S., "A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands," in *Proc. of 4th Int. Conf. on Spoken Language Processing (ICSLP 1996)*, Philadelphia, PA, USA, Oct. 1996, pp. 426–429.

[4] Hermansky, H.; Tibrewala, S.; Pavel, M., "Towards ASR on Partially Corrupted Speech," in *Proc. of 4th Int. Conf. on Spoken Language (ICSLP 1996)*, Philadelphia, PA, USA, Oct. 1996, pp. 462–465.

[5] Sumby, W. H.; Pollack, I., "Visual Contribution to Speech Intelligibility in Noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, Mar. 1954.

[6] Stork, D. G.; Hennecke, M. E.; Prasad, K. V., "Visionary Speech: Looking Ahead to Practical Speechreading Systems," in *Speechreading by Humans and Machines*, Stork, D. G.; Hennecke, M. E., Ed. Springer, Berlin, Germany, 1996.

[7] Neti, C.; Potamianos, G.; Luettin, J.; Matthews, I.; Glotin, H.; Vergyri, D.; Sison, J.; Mashari, A.; Zhou, J., "Audio-Visual Speech Recognition," Tech. Rep., Center Lang. Speech Process., Johns Hopkins University, Baltimore, MD, USA, 2000.

[8] Jain, U.; Siegler, M. A.; Doh, S.-J.; Gouvea, E.; Huerta, J.; Moreno, P. J.; Raj, B.; Stern, R. M., "Recognition of Continuous Broadcast News with Multiple Unknown Speakers and Environments," in *Proc. of ARPA Speech Recognition Workshop*, Harriman, NY, USA, Feb. 1996, pp. 61–66.

[9] Ming, J.; Hanna, P.; Stewart, D.; Owens, M.; Smith, F. J., "Improving Speech Recognition Performance by Using Multi-Model Approaches," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1999)*, Phoenix, AZ, USA, Mar. 1999, pp. 161–164.

[10] Kittler, J.; Hatef, M.; Duin, R.; Matas, J., "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[11] Kamel, M. S.; Wanas, N. M., "Data Dependence in Combining Classifiers," in *Multiple Classifier Systems*, Windeatt, T.; Roli, F., Ed., pp. 1–14. Springer, Berlin, Germany, Jun. 2003.

[12] Garg, A.; Potamianos, G.; Neti, C.; Huang, T. S., "Frame-Dependent Multi-Stream Reliability Indicators for Audio-Visual Speech Recognition," in *Proc. of Int. Conf. Multimedia and Expo (ICME 2003)*, Baltimore, MD, USA, Jul. 2003, pp. 605–608.

[13] Varga, P.; Moore, R. K., "Hidden Markov Model Decomposition of Speech and Noise," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1990)*, Albuquerque, NM, USA, Apr. 1990, pp. 845–848.

[14] Nefian, A. V.; Liang, L.; Pi, X.; Liu X.; Murphy, K., "Dynamic Bayesian Networks for Audio-Visual Speech Recognition," *EURASIP Journal on Applied Signal Processing*, vol. 11, no. 1, pp. 1274–1288, Jan. 2002.

[15] Kolossa, D.; Zeiler, S.; Vorwerk, A.; Orglmeister, R., "Audio-visual Speech Recognition with Missing or Unreliable Data," in *Proc. of Int. Conf. on Auditory-Visual Speech Processing (AVSP 2009)*, Norwich, UK, Sept. 2009, pp. 117–122.

[16] Fiscus, J. G., "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU 1997)*, Santa Barbara, CA, USA, Dec. 1997, pp. 347–352.

[17] Shivappa, S. T.; Rao, B. D.; Trivedi, M. M., "An Iterative Decoding Algorithm for Fusion of Multimodal Information," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 1–10, Nov. 2007.

[18] Shivappa, S. T.; Rao, B. D.; Trivedi, M. M., "Multimodal Information Fusion Using the Iterative Decoding Algorithm and its Application to Audio-Visual Speech Recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, Las Vegas, NV, USA, Mar. 2008, pp. 2241–2244.

[19] Scheler, D.; Walz, S.; Fingscheidt, T., "On Iterative Exchange of Soft State Information in Two-Channel Automatic Speech Recognition," in *Proc. of 10th ITG Conf. on Speech Communication*, Braunschweig, Germany, Sept. 2012, pp. 55–58.

[20] Shafi, M.; Chung, P. W. H., "A Hybrid Method for Eyes Detection in Facial Images," in *Proc. of Int. Conf. on Computer Science (WASET 2008)*, Singapore, Aug. 2008, pp. 99–104.

[21] Guitarte, J. F.; Lukas, K.; Frangi, A. F., "Low Resource Lip Finding and Tracking Algorithm for Embedded Devices," in *Proc. of Int. Conf. on Audio-Visual Speech Processing (AVSP 2003)*, St. Jorioz, France, Sep. 2003, pp. 111–116.

[22] Hsu, R.-L.; Abdel-Mottaleb, M.; Jain, A. K., "Face Detection in Color Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696–705, May 2002.

[23] Eveno, N.; Caplier, A.; Coulon, P.-Y., "A New Color Transformation for Lips Segmentation," in *Proc. of 4th Workshop on Multimedia Signal Processing*, Cannes, France, Oct. 2001, pp. 3–8.

[24] Cootes, T. F.; Taylor, C. J., "Statistical Models of Appearance for Computer Vision," Tech. Rep., Institute of Population Health, University of Manchester, Manchester, UK, Oct. 2001, http://www.isbe.man.ac.uk/.

[25] ETSI, *ETSI ES 202 050 V1.1.5 Advanced Front-End Feature Extraction Algorithm*, European Telecommunication Standards Institute, Jan. 2007.

[26] Cooke, M.; Barker, J.; Cunningham, S.; Shao, X., "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, Nov. 2006.

[27] ITU-T, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Dec. 2011.