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

Dynamic Stream Weighting for Turbo-Decoding-Based Audiovisual ASR

Sebastian Gergen[1], *Steffen Zeiler*[1], *Ahmed Hussen Abdelaziz*[2], *Robert Nickel*[3], *Dorothea Kolossa*[1]

[1] Cognitive Signal Processing Group, Institute of Communication Acoustics, Ruhr-University Bochum
[2] International Computer Science Institute, Berkeley
[3] Department of Electrical and Computer Engineering, Bucknell University

sebastian.gergen@rub.de, steffen.zeiler@rub.de, ahmed.hussenabdelaziz@rub.de,
robert.nickel@bucknell.edu, dorothea.kolossa@rub.de

Abstract

Automatic speech recognition (ASR) enables very intuitive human-machine interaction. However, signal degradations due to reverberation or noise reduce the accuracy of audio-based recognition. The introduction of a second signal stream that is not affected by degradations in the audio domain (e.g., a video stream) increases the robustness of ASR against degradations in the original domain. Here, depending on the signal quality of audio and video at each point in time, a dynamic weighting of both streams can optimize the recognition performance. In this work, we introduce a strategy for estimating optimal weights for the audio and video streams in turbo-decoding-based ASR using a discriminative cost function. The results show that turbo decoding with this maximally discriminative dynamic weighting of information yields higher recognition accuracy than turbo-decoding-based recognition with fixed stream weights or optimally dynamically weighted audiovisual decoding using coupled hidden Markov models.

Index Terms: Audiovisual speech recognition, Turbo decoding, Stream weighting

1. Introduction

Automatic speech recognition (ASR) is essential in situations in which a hands-free interaction between a human and a machine is desired. Due to the many technical advances that were made in the last few decades, modern ASR systems have become quite accurate and reliable when operated in favorable environmental conditions. The performance drops considerably, however, when noise and/or reverberation degrade the incoming signal. An incorporation of information from a different (e.g. non-acoustic) modality, which is not affected by the degradations in the original acoustic modality, helps to increase the noise robustness and the recognition performance. It has been shown that with a combination of audio and video data it is possible to increase speech recognition accuracy when compared to ASR systems that employ audio data streams only (e.g., [1, 2]). Currently, the main techniques used for the integrated decoding of audio and video streams are coupled HMMs (CHMM, [3]), turbo-decoding (TD, [4]), and approaches based on deep learning [5, 6, 7]. In dealing with the integration it is important to control the contribution of each stream within the decoding procedure in order to ensure an optimal performance. Fixed stream weights (FSWs) for a given data-set were analyzed in [1, 8] and time-adaptive weights (dynamic stream weights, DSW) were proposed in [9]. In the latter paper, an algorithm to estimate DSWs for CHMM-based ASR was introduced for different signal-to-noise ratios (SNR) in the audio stream. Unfortunately, the method proposed in [9] for CHMM-based ASR does not readily carry over to TD-based ASR.

In this contribution, we introduce a new estimation strategy for optimal DSWs for TD-based ASR. We propose to use the stream weights of audio and video streams to maximize a discriminative cost function in each time frame and TD-iteration. This cost function uses the knowledge of the correct word sequence of the current sentence in form of an *oracle* path through the HMM-states as well as the N-best (*confusion*) paths which result in wrong word sequences. With these oracle-based DSWs, estimation algorithms for blind estimation of high quality DSWs will be developed in the future.

The paper is structured as follows. In Section 2 we introduce the basic concepts of audiovisual ASR with CHMM and TD and we discuss the estimation of optimal SNR- and noise-type-dependent fixed stream weights. In Section 3, we describe the proposed estimation method for dynamic stream weights for TD. Our experimental setup is presented in Section 4. The accuracy of ASR systems with and without the proposed method is evaluated in Section 5.

2. Background

It is helpful to briefly review the main concepts behind CHMMs and TD for audiovisual fusion and to define the associated mathematical notation. A broader and more comprehensive introduction to the field of speech recognition with CHMMs and turbo decoding is found in [10], [11], and [4] for example.

When audiovisual data is used for ASR, a fusion strategy is required that jointly evaluates data from both modalities. For this purpose, a comprehensive initial study by Nefian et al. [1] had found the coupled HMM to be the optimal method among a range of graphical-model-based approaches. However, in more recent work, turbo decoding has proven to be even more effective at integrating both modalities [12]. The two methods are described in the following two subsections 2.1 and 2.2.

2.1. Coupled HMM Decoding

Coupled HMM decoding is a commonly applied technique to fuse information from an audio stream with that of a visual stream. With coupled HMMs it is readily possible to model the temporal asynchronicities that typically occur between video and audio data. A more comprehensive discussion of CHMMs and their advantages can be found in [1] and [2]. The joint au-

This project was supported by the German research foundation DFG (project KO3434/4-1).

http://dx.doi.org/10.21437/Interspeech.2016-166

diovisual observation likelihood of a CHMM is computed as

$$p(o_a, o_v|q_a, q_v) = b_a(o_a|q_a)^{\lambda_c} \cdot b_v(o_v|q_v)^{1-\lambda_c}, \quad (1)$$

where b_a and b_v are the individual audio and video observation likelihoods of the corresponding single-modalitiy observation o and state q. The term λ_c weights the contribution of the modalities to the joint likelihood, and may be fixed or varying over time.

2.2. Turbo Decoding

Originally, turbo decoding has been developed for the purpose of convolutional error correction and channel decoding in digital transmission systems [13, 14]. Recently, TD was introduced into the field of ASR as means to solve the data fusion problem in multi-modal recognition tasks [15, 4, 16]. In the iterative process of TD (see Fig. 1), information which is extracted from state posteriors is exchanged between different decoders. To iteratively improve audio and video likelihoods, the observation likelihoods b_a and b_v are modified by g_a and g_v as

$$\tilde{b}_a(o_a|q_a) = b_a(o_a|q_a) \cdot g_a(q_a)^{\lambda_p \lambda_v}, \quad (2)$$

$$\text{and} \qquad \tilde{b}_v(o_v|q_v) = b_v(o_v|q_v) \cdot g_v(q_v)^{\lambda_p \lambda_a}. \quad (3)$$

In both equations, λ_p can be interpreted as a constant weighting exponent, which balances the influence of the prior probability versus the likelihood of the previous iteration. Exponents λ_a and λ_v can be used to control the weight of the likelihood modification. We may either chose λ_a and λ_v to be fixed for a given expected SNR value or we may adjust them dynamically over time according to some type of estimated fidelity criterion.

Using the forward-backward algorithm, in each half-iteration we obtain new state posteriors $\tilde{\gamma}$, which contain information about the likelihood, the prior probability and the extrinsic probability [4]. This extrinsic probability, which is passed on to the other decoder for the next half-iteration, is computed for state q at time frame t by removing all excess information, i.e.

$$\dot{\gamma}(q_t) \propto \frac{\tilde{\gamma}(q_t)}{b(o_t|q_t) \cdot g(q_t)}. \quad (4)$$

For L video states we define $\dot{\gamma}_v = [\,\dot{\gamma}_v(1) \ \ldots \ \dot{\gamma}_v(L)\,]^{\mathrm{T}}$. Analogously, we define vectors $\dot{\gamma}_a$, g_a, and g_v. To transfer the extrinsic probabilities $\dot{\gamma}_v$ and $\dot{\gamma}_a$ from one modality in the other, we are mapping the $\dot{\gamma}_v$ and $\dot{\gamma}_a$ vectors to the g_a and g_v vectors via a linear transformation

$$g_a = T_{va}\dot{\gamma}_v, \quad (5)$$

$$\text{and} \qquad g_v = T_{av}\dot{\gamma}_a. \quad (6)$$

The matrices T_{va} and T_{av} are derived from the relationships between the state spaces for the video and the audio streams respectively (see [4] for example). In our case, three states per phoneme are used in the acoustic model and one state per phoneme in the video model, which, together with the fact that we associate acoustic and phonetic states according to the phoneme identity, defines the structure of these matrices.

2.3. Fixed Stream Weights

Fixed stream weights are designed to provide a noise-scenario-dependent weighting for audio and video data, with a *noise scenario* being defined by the type and level of noise. The FSWs are fixed for all signals of the specific noise scenario and over all possible iterations of decoding. Thus, the FSWs can be applied when a noise estimation algorithm provides information about the acoustic scenario.

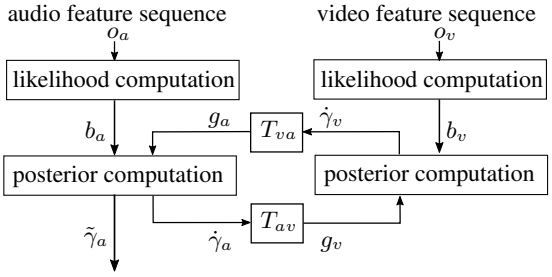


Figure 1: Schematic overview of turbo-decoding for audio-visual ASR. The left column comprises a forward-backward-algorithm-based audio-only ASR system. For TD, a second modality (video) is added and extrinsic probabilities $\dot{\gamma}_a$ and $\dot{\gamma}_v$ are exchanged between decoders. After a predefined number of iterations, a best path search through the audio posteriors $\tilde{\gamma}_a$ reveals the final best word sequence.

To estimate optimal FSWs, we follow the strategy that is introduced in [17]. A grid search is performed to estimate optimal weighting coefficients of audio and video data in different noise conditions. For the computation of the joint audiovisual state likelihood for CHMM-decoding (Eq. (1)), the grid search has to cover only the single parameter which we define for the grid search as $\lambda_c^{\text{FSW}} \in \{0.1, 0.2, \ldots, 0.9\}$. For the computation of the modified audio and video likelihoods for TD (Eq. (2) and (3)), a grid search over the two parameters λ_m with $m \in \{a, v\}$ is required. A low (high) weighting of one modality does not necessarily result in a high (low) weighting of the other, which is a conceptual difference between the weighting of streams in a CHMM and in TD. Furthermore, the admissible range for λ_m for TD is not limited to values between 0 and 1.

3. Dynamic Stream Weights for Turbo-Decoding

Dynamic stream weights are designed to optimally weigh the information provided by the audio and video streams at each time frame. In [9, 17], an Expectation-Maximization (EM) based strategy is introduced which dynamically estimates the λ_c^{DSW} values for Eq. (1).

Unfortunately, the algorithms proposed in [9, 17] do not readily carry over from CHMM-based to TD-based stream weighting. For CHMMs, the stream weight is used to balance the likelihoods of both modalities. The audio and video weights sum to 1 and thus only the single parameter λ_c is needed (Eq. (1)). For DSW in TD, however, the stream weight balances the state likelihood in one modality with prior information from the other. One has to estimate two parameters λ_a and λ_v which do not need to complement to 1 at all (see Eq. (2) and (3)). Furthermore, these parameters may vary within the iterative process of the TD itself.

We introduce a new strategy here for the estimation of DSWs, which incorporates principles from discriminative training, and we apply it to turbo-decoding-based audiovisual ASR. The goal is to maximize (at each time frame t of each TD-iteration) the modified likelihood of the state of the most likely sequence which would result in the correctly decoded word sequence (i.e., a state from an *oracle* path $\boldsymbol{s}^{\text{orac}}$). At the same time we strive to reduce the modified likelihoods of the N most likely different sequences of states that would result in incorrectly decoded word sequences (i.e., the *confusion* paths $\boldsymbol{s}^{\text{conf}}$).

3.1. *Oracle* and *confusion* paths search

The cost function in the estimation of DSWs involves two sets of paths: N *confusion* paths $\boldsymbol{s}^{\text{conf}}$ and one *oracle* path $\boldsymbol{s}^{\text{orac}}$ for audio and video modalities $m \in \{a, v\}$. The *confusion* paths arise as solutions to the best-path search through the matrix of state posteriors of a composite HMM representing all possible sentences for which we use the grammar of our employed GRID database [18]. The composite HMM is constructed according to the task grammar *[command,color,preposition,letter,digit,adverb]* by the parallel and serial connection of all single-word HMMs for each respective word type. We define the confusion path set

$$\boldsymbol{s}_{m,n}^{\text{conf}} = [q_{m,n}^{\text{conf}}(1), q_{m,n}^{\text{conf}}(2), \ldots, q_{m,n}^{\text{conf}}(T)] \text{ ,with } 1 \leq n \leq N$$

as those N-best Viterbi paths through the matrix of state posteriors of the composite HMM for a given utterance corresponding to different incorrect word sequences (with time frames $t = 1, \ldots, T$). Note that $N + 1$ best paths have to be evaluated in the path search in order to ensure to obtain N *confusion* paths; a path within the $N + 1$ best paths that results in the *correct* word sequence can be discarded.

The oracle path s^{orac} is the path through the state posteriors that carries the highest likelihood of all word sequences belonging to the correct word sequence. It is found by first constructing a forced-alignment HMM by sequential concatenation of word HMMs according to the sentence label. Afterwards we find the single best Viterbi path through the state posteriors of this HMM. To find $\boldsymbol{s}_m^{\text{orac}} = [q_m^{\text{orac}}(1), q_m^{\text{orac}}(2), \ldots, q_m^{\text{orac}}(T)]$, we translate the state indices of the path through the forced-alignment HMM for a single word sequence into the state indices of the composite HMM describing all possible word sequences.

3.2. Discriminative cost function

The boosted mutual information criterion was defined in [19] and [20] as a cost function for parameter estimation within discriminative model training [21]. In this work we adapt this criterion and define a cost function for modality m at each time frame $t = 1, \ldots, T$ (note, that we neglect the time frame index in the following equation) as

$$J(\lambda_m) = \log \frac{\tilde{b}_m(o_m|q_m^{\text{orac}})^\kappa}{\sum_{n=1}^{N}(\tilde{b}_{m,n}(o_m|q_{m,n}^{\text{conf}})^\kappa \exp(-\alpha \cdot \delta_{m,n}))} \tag{7}$$

where α and κ are constant weighting coefficients. $\delta_{m,n}$ is a distance measure between the *oracle* and the n-th *confusion* path. This distance measure penalizes deviations of a *confusion* path to the *oracle* path (low penalty for deviations towards states corresponding to the correct word and a higher penalty for states of different words). By finding the optimum of the cost function, we obtain the stream weight that maximizes the ratio of the modified likelihood for a state of the *oracle* path to the sum of the modified likelihoods of the states in the *confusion* paths. We evaluate the cost function at each time frame and in each TD iteration for both of the modalities for a range of λ values, and select that λ for which (7) becomes largest.

4. Experimental Setup

For our experiments we used a subset of the audiovisual data provided in the GRID database [18]. The GRID database contains 34000 small-vocabulary semantically unpredictable sentences with fixed grammar uttered by 34 speakers.

We divided the dataset into a training set (90% of the data) for speaker-dependent model training and a development set (5% of the data) for the stream weight estimation. Since we are working with oracle label information per sentence, all recognition results in this contribution (for FSW and DSW) are necessarily based on this development set. For the next phase of development, a third subset, the test set (5% of the data), will be used to evaluate the speech recognition performance in conjunction with an automatic estimation of fixed as well as dynamic stream weights as in [17], which is under development.

The following analyses were done at different levels of fidelity of the underlying audio signals, i.e. with clean speech as well as speech in babble noise and white noise. For the latter cases, signals at different SNR levels were created (SNR $= \{0, 5, 10, 15\}$ dB). The additive noise was taken from the NOISEX-92 database [22]. For the video streams, the original clean signals from the GRID database were used.

The (MFCC-based) audio features and video features (DCT features of the mouth region) that we extracted were the same as the ones described in [17]. Here, however, we used a sampling rate of $f_s = 16$ kHz instead of 8 kHz for the audio signals and we kept the full dimension of 39 for the audio features (while the dimension was reduced in [17]). The dimension of the video features was reduced to 31 as in [17] by means of a linear discriminant analysis. The speaker-dependent models for the ASR system were 51 single-stream left-to-right word HMMs and one additional HMM for silence (384 states for audio and 128 states for video). All experiments were carried out with our Java Audiovisual SPEech Recognizer JASPER[3].

Within the ASR system, we used CHMM-based decoding and TD. In the TD procedure we used a flat prior $g_a(q_a) = 1, \ \forall\ q_a$ for the audio state posterior calculation in the first of in total 4 TD-iterations and then performed a best-path search for the decoding based on the audio posteriors. We selected a constant weight for the prior probability $\lambda_p = 0.001$.

To cover a wide range of values with a rather small number of grid-search iterations for the estimation of FSWs (on the development set), we defined an exponential search space such that $\lambda_{a/v}^{\text{FSW}} \in \{2^0, 2^1, ..., 2^7\}$. The parameters of the cost function for the DSW estimation were set to $\alpha = 0.5$ and $\kappa = 0.75$ and we evaluated the cost function for 1000 values of λ, with $10^{-3} \leq \lambda \leq \frac{1}{\lambda_p}$, and evaluated $N = 15$ confusion paths. The performance of the various ASR systems was measured with the recognition accuracy as defined in [17].

5. Speech Recognition Results

Table 1 presents the resulting FSWs for CHMM and TD-based data fusion. For both decoding approaches, the amount of video information that leads to a good ASR performance increases with a decrease in the audio-SNR. This is reflected in a decreasing trend of λ_c over the SNRs for CHMM decoding and a decreasing trend of the ratio λ_a/λ_v over the SNRs for TD.

Table 2 shows the accuracy of audiovisual ASR via CHMM decoding. These results serve as a reference for the TD-based recognition. They are in agreement with the results reported in [17]. For purely audio-based decoding, we observe that for clean signals, the ASR performs very well, but with an increasing amount of noise, the ASR accuracy drops considerably. The accuracy of ASR based on pure video information is 86.40%. Combining audio and video information with equal weighting ($\lambda_c = 0.5$) leads to an improvement in ASR performance (compared to audio only) over all SNR levels, but especially for low-SNR scenarios. However, in one noise scenario (white noise, 0

Table 1: Oracle-based fixed stream weighting for different noise scenarios.

Noise Type	SNR [dB]	CHMM λ_c^{FSW}	Turbo decoding λ_a^{FSW}	λ_v^{FSW}	$\lambda_a^{\text{FSW}}/\lambda_v^{\text{FSW}}$
Clean	-	0.8	32	8	4.00
Babble	15	0.7	128	32	4.00
	10	0.6	32	32	1.00
	5	0.5	32	32	1.00
	0	0.3	8	128	0.06
White	15	0.6	32	32	1.00
	10	0.5	8	32	0.25
	5	0.3	8	32	0.25
	0	0.2	8	128	0.06

Table 2: ASR accuracy (in %) of CHMM-based decoding.

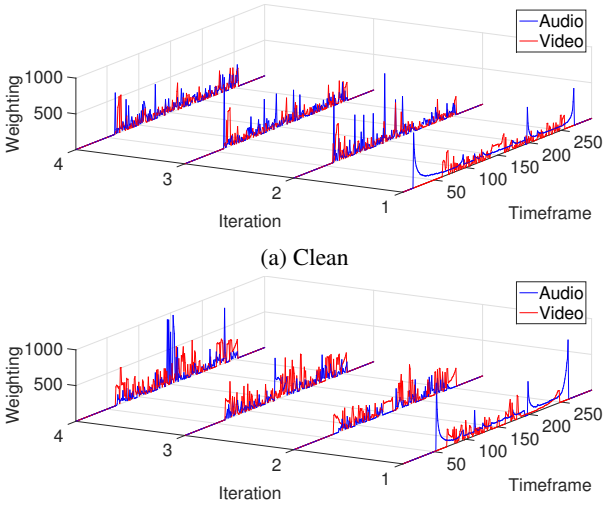
Noise Type	SNR [dB]	Audio only $\lambda_c = 1$	Audio-visual $\lambda_c = 0.5$	λ_c^{FSW}	λ_c^{DSW}
Clean	-	97.83	98.04	98.67	99.04
Babble	15	90.64	95.75	96.37	97.39
	10	78.85	93.79	94.04	95.82
	5	55.37	91.07	91.07	93.79
	0	31.43	86.89	88.44	90.91
White	15	77.87	94.11	94.25	95.99
	10	54.43	91.45	91.45	93.86
	5	32.71	88.21	89.10	91.41
	0	21.06	83.94	87.56	89.85
Average	-	60.02	91.47	92.33	94.23

Table 3: ASR accuracy (in %) using turbo decoding and different stream weights $\boldsymbol{\lambda} = [\lambda_a, \lambda_v]$.

Noise Type	SNR [dB]	$\boldsymbol{\lambda} =$ $[32, 32]$	$\boldsymbol{\lambda} =$ $[\lambda_a^{\text{FSW}}, \lambda_v^{\text{FSW}}]$	$\boldsymbol{\lambda} =$ $[\lambda_a^{\text{DSW}}, \lambda_v^{\text{DSW}}]$
Clean	-	97.20	98.81	99.46
Babble	15	96.80	97.02	98.71
	10	95.17	95.17	97.56
	5	92.32	92.32	95.83
	0	85.16	88.43	92.81
White	15	95.81	95.81	97.83
	10	93.40	93.43	96.36
	5	89.16	90.19	94.39
	0	80.08	87.92	91.59
Average	-	91.68	93.22	96.06

dB) the performance of coupled-HMM-based audio-visual ASR using equal stream weights is worse than the video-only ASR. Improvements in ASR performance can be achieved by the application of the optimally estimated FSWs: then, bimodal decoding outperforms both of the unimodal approaches in all scenarios. When the oracle DSWs are used, the ASR performance is improved again in all noise types and levels.

The ASR accuracy for TD-based ASR is presented in Table 3. We employ the vector $\boldsymbol{\lambda} = [\lambda_a, \lambda_v]$ to represent both weights in a compact form. TD-based ASR performs better than the CHMM-based one in many but not all noise scenarios for equally weighted information from both modalities. Note that we selected the equal weighting $\boldsymbol{\lambda} = [32, 32]$ as this weighting provided the best average accuracy (over all noise scenarios) of all equal weightings. When using optimal FSWs our TD-based ASR outperforms the respective CHMM-based ASR for all but one case and the average performance of TD-based ASR using fixed weights is between the performance of FSW- and DSW-based CHMM ASR. The newly proposed dynamic weighting of audio and video streams for TD-based ASR clearly results in the highest accuracy in our experiments. Especially for low-SNR scenarios the accuracy increases considerably. Compared to the case of fixed stream weights, the average performance



(a) Clean

(b) Babble noise, SNR: 5dB

Figure 2: Dynamic weights of audio and video streams in four TD-iterations for the GRID-sentence *'lay green at s one please'* with no additional noise (a), or in the presence of babble noise (b).

increases from 93.22% to 96.06%.

An example of oracle dynamic stream weights for one sentence is shown in Fig. 2 for two of the acoustic scenarios. Especially in the second TD-iteration the difference of the audio weighting for the clean signal compared to the noisy scenario is observable. For both scenarios we observe a similar weighting of audio in the very first TD-iteration which is related to the initialization with a flat prior. The overall weighting fluctuates over time and over the TD-iterations. However, it is observable that the video information is assigned a higher weighting for the noisy case (Fig. 2b) than in the case of noise-free audio signals (Fig. 2a).

6. Conclusions

The integration of audio and video data for speech recognition significantly improves ASR results in noisy scenarios. The best performance is typically achieved by using turbo-decoding for this purpose. In this paper, we have shown how a dynamic weighting of audio and video information can notably benefit such turbo-decoding-based speech recognition.

In order to obtain optimal stream weights based on label information, we have introduced a discriminative cost function that has allowed us to find maximally discriminative dynamic stream weights. The results have shown that the speech recognition accuracy can be improved notably, both in comparison to fixed stream weights and in comparison to dynamically weighted CHMM-decoding.

However, the computation of these oracle DSWs relies on *oracle* knowledge about the correct word sequence, and the method is hence not applicable directly to ASR. Instead, the oracle stream weights can serve as high-quality training targets for deep-neural-network or regression stream weight estimators, which only rely on signal- and classification-quality criteria to find suitable DSWs. Hence, the proposed method serves as an intermediary step, providing us with arbitrarily large sets of maximally discriminative training targets for such stream-weight estimators, the training and application of which is the goal of our ongoing work.

7. References

[1] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1–15, 2002.

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep 2003.

[3] A. Vorwerk, S. Zeiler, D. Kolossa, R. F. Astudillo, and D. Lerch, "Use of missing and unreliable data for audiovisual speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data*, D. Kolossa and R. Haeb-Umbach, Eds. Springer, 2011, pp. 345–375. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-21317-5

[4] S. Receveur, D. Scheler, and T. Fingscheidt, "A turbo-decoding weighted forward-backward algorithm for multimodal speech recognition," in *5th International Workshop on Spoken Dialog Systems*, 2014, pp. 4–15.

[5] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 1–8.

[6] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda, and S. Hayamizu, "Audio-visual speech recognition using deep bottleneck features and high-performance lipreading," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015, pp. 575–582.

[7] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2130–2134.

[8] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000. [Online]. Available: http://dx.doi.org/10.1109/6046.865479

[9] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "A new EM estimation of dynamic stream weights for coupled-HMM-based audio-visual ASR," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1527–1531.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[11] R. Haeb-Umbach and D. Kolossa, *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, D. Kolossa and R. Haeb-Umbach, Eds. Springer, 2011.

[12] S. Receveur, R. Weiss, and T. Fingscheidt, "Turbo automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 99, pp. 1–1, 2016.

[13] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo codes," in *IEEE International Conference on Communications*, 1993, pp. 1064–1070.

[14] C. Berrou and A. Glavieux, "Near optimum error-correcting coding and decoding: Turbo Codes," *IEEE Transactions on Communications*, vol. 44, no. 10, pp. 1261–1271, Oct. 1996.

[15] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 2241–2244.

[16] S. Zeiler, R. Nickel, N. Ma, G. J. Brown, and D. Kolossa, "Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 1–2.

[17] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 23, no. 5, pp. 863–876, 2015.

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

[19] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1986, pp. 999–999.

[20] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 4057–4060.

[21] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1060–1089, 2013.

[22] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG-10 noise database," in *Technical Report IZF 1988-3, TNO Institute for Perception, Soesterberg, The Netherlands*, 1988.