On Likelihood Histogram Equalization for Multimodal Automatic Speech Recognition

Simon Receveur and Tim Fingscheidt
Institute for Communications Technology, Technische Universität Braunschweig,
38106 Braunschweig, Deutschland, Email: {s.receveur, t.fingscheidt}@tu-bs.de

Introduction

The incorporation of visual information has been shown as effective approach to robust automatic speech recognition (ASR) [1]. In the past a number of techniques have been presented addressing the audio-visual information fusion in broadly two manners: feature fusion and decision fusion approaches [2].

Among the decision fusion frameworks, the coupled hidden Markov model (CHMM) permits asynchrony between the respective audio and visual states within the word boundaries, while retaining the natural dependency of the audio and video input feature vectors [3]. In particular, the information fusion problem is commonly solved by applying a weighted product rule, where the joint probability distribution of the observation likelihoods is composed by the product of the individual observation likelihoods with respective exponents [4]. Within this fusion process, the relative influence of the information sources or so-called *streams* is controlled by these exponents, also called weighting parameters or *stream weights*. For audio-visual ASR, the audio stream weight is typically chosen dependent on the signal-to-noise ratio (SNR) [5].

This paper is organized as follows: After shortly reviewing the CHMM, we examine numerical issues of the respective observation likelihoods and compensate numerical mismatches of the streams by employing a histogram equalization (EQ) technique. Applied to an audio-visual speech recognition task, we show the sensitivity of the recognition results at different SNRs. The paper ends with some concluding remarks.

Coupled HMM (CHMM)

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 $\lambda = \{\boldsymbol{\pi}; \mathbf{A}; \mathcal{B}\}$, 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 $i \in \mathcal{S} = \{1, \ldots, N\}$, $\mathbf{A} = \{a_{j,i}\}_{j,i \in \mathcal{S}}$, the matrix of state transition probabilities $a_{j,i} = \mathrm{P}(s_t = i \,|\, s_{t-1} = j)$, and $\mathcal{B} = \{b_i(\mathbf{x}_t)\}_{i \in \mathcal{S}}$, the set of d_o-variate emission probability density functions (pdfs) $b_i(\mathbf{x}_t) = \mathrm{p}(\mathbf{x}_t \,|\, s_t = i)$. Note that we use $\mathrm{P}(\cdot)$ for probabilities and $\mathrm{p}(\cdot)$ for pdfs (or their values). Moreover, we consider a further observation stream $\mathbf{y}_1^T = \mathbf{y}_1, \ldots, \mathbf{y}_T$ originating from a different sensor or modality; its feature vectors take on values $\mathbf{y}_t = \mathbf{u}_t \in \mathbb{R}^{d_u}$ from a different feature space $\mathbb{R}^{d_u}$ with vector dimension d_u. Besides, we assume two state-level *maximum-a-posteriori* (MAP) recognizers processing one of the given observation feature streams, respectively. Each recognizer applies an individual hidden Markov model (HMM) trained to match the handled observations, one in state space $\mathcal{S}$, the other one in state space $\mathcal{R}$.

CHMMs are capable of modeling the inherent asynchrony between the audio and video channels. In particular, the hidden states of either HMM are allowed to interact, but retain their own observations at the same time [3, 5]. Thus, the coupled stationary state transition probability

$$\mathbf{A}^{(s),(r)} = \{a_{j,i} \cdot a_{\ell,k}\}_{j,i \in \mathcal{S},\, \ell,k \in \mathcal{R}}, \tag{1}$$

as well as the coupled emission

$$\begin{aligned} b_{i,k}^{(s),(r)}(\mathbf{o}_t, \mathbf{u}_t) &= b_i(\mathbf{o}_t)^{\varphi_o} \cdot b_k(\mathbf{u}_t)^{\varphi_u} \\ &= \mathrm{p}(\mathbf{o}_t \,|\, s_t = i)^{\varphi_o} \cdot \mathrm{p}(\mathbf{u}_t \,|\, r_t = k)^{\varphi_u}, \\ &\qquad \forall\, (i,k) \in \mathcal{S} \times \mathcal{R}, \end{aligned} \tag{2}$$

can both be combined from the two marginal unimodal HMMs trained on an audio-visual training corpus, respectively. Note that in the CHMM system we utilize exponential stream weights φ_o and φ_u on the audio and video emissions, respectively. These weights are separately optimized during training — φ_o as SNR-dependent —, letting $0 \leq \varphi_o, \varphi_u \leq 1$ and $\varphi_o + \varphi_u = 1$.

Histogram Equalization (EQ)

As a basis for computing the joint probability distribution of the observation likelihoods, we extracted shaped-based features of order 11 for each speaker at the visual frontend [6], respectively. As acoustic features we employed 13 MFCC coefficients, 1st- and 2nd-order derivatives and an additional log-energy parameter.

The two emissions $b_i(\mathbf{o}_t)$ and $b_k(\mathbf{u}_t)$ in (2) complement each other and ideally, the correct information defeats the incorrect one. However, the numerical prerequisites for this to happen are rarely fulfilled in audio-visual ASR: First of all, we found that the emissions of both streams differ widely with respect to number range and statistical properties. In particular, even on a logarithmic scale the mean of the video observation likelihoods is much smaller than the mean of the audio observation likelihoods (-256.1 vs. -41.9), whereas on the other hand the standard deviation of the video observation likelihoods is considerably larger than the standard deviation of the other (223.4 vs. 28.9). Due to the lower mean, the video stream is principally underestimated within the weighted product rule (2).

A simple yet effective relief to this underestimation is given by means of an EQ: During training, the means μ_o, μ_u and standard deviations σ_o, σ_u of the respective emissions $b_i(\mathbf{o}_t), b_k(\mathbf{u}_t)$ are estimated. Prior to recognition, the emissions of the second stream are equalized to match the histogram of the first:

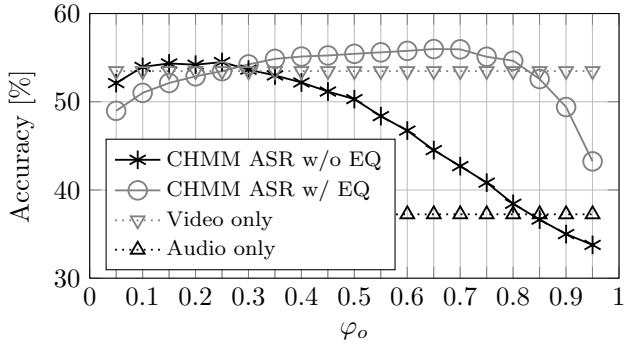

Figure 1: Word accuracy (% ACC) vs. variation of the exponential stream weight φ_o at an SNR of 0 dB.

CHMM	Word Accuracy [%]			
	0 dB	5 dB	10 dB	15 dB
w/o EQ	54.5	65.9	76.0	84.3
w/ EQ	56.0	66.5	76.6	84.2

(a) Optimal CHMM recognition results in word accuracy (% ACC) for various SNRs.

CHMM	$\varphi_{o,opt}$			
	0 dB	5 dB	10 dB	15 dB
w/o EQ	0.25	0.35	0.75	0.80
w/ EQ	0.65	0.85	0.90	0.95

(b) Optimal stream weights $\varphi_{o,\mathrm{opt}}$ for various SNRs.

Table 1: Table of the optimal CHMM recognition results and exponential stream weights φ_o for various SNRs.

$$\bar{b}_k^{(r)}(\mathbf{u}_t) = \left(b_k^{(r)}(\mathbf{u}_t) - \mu_u\right) \cdot \frac{\sigma_o}{\sigma_u} + \mu_o. \qquad (3)$$

Evaluation

To examine the presented EQ technique, we selected 20 (10 male and 10 female) speakers of the GRID audio-visual speech corpus [7] containing audio and video recordings of 1000 utterances per speaker. Moreover, based on ITU-T P.56 the audio recordings were additionally interfered with white noise at various signal-to-noise ratios (SNRs). We trained separate HMM sets for each MAP recognizer (either video, or undisturbed audio) on 800 training sentences from the same speakers. The remaining 200 (disjoint) utterances were employed in the test set. Each HMM set was composed of 51 word HMMs corresponding to the GRID corpus overall vocabulary. The word HMMs had a linear topology and four states per phoneme, respectively, whose state emission pdfs were modeled with Gaussian mixture models (GMMs) of order 4 and diagonal covariance matrices.

Figure 1 shows the recognition results in word accuracy (% ACC) vs. variation of the exponential stream weight φ_o at an SNR of 0 dB. The dotted lines with triangular markers represent single-channel baselines ($\triangle$: audio, $\triangledown$: video). The lines with ($\circ$) and ($*$) markers illustrate an audio-visual CHMM approach, with or without applying an EQ, respectively. The following single-modality accuracies were achieved: 53.5 % on the video-only test corpus, while the audio-only results vary from 37.2 % at 0 dB to 91.3 % at 30 dB. Table 1a indicates that the use of an EQ technique is particularly suitable at low SNRs (up to 15 dB), where the video-only results surpass the audio-only results. At an SNR of 0 dB, the use of an EQ increases the influence of the video emissions within the joint probability distribution (2) outperforming the conventional CHMM approach w/o EQ by about 1.5 % absolute. Moreover, the relatively flat curve of the CHMM approach w/ EQ indicates an increased robustness against a variation of the exponential stream weight φ_o. With higher SNR and thus better audio-only results, however, the inherent underestimation of the video emissions without any EQ may be even desired indicated by the better overall CHMM results and the high values of $\varphi_{o,\mathrm{opt}}$ in Table 1b.

Conclusions

In this contribution we examined numerical issues of audio and video observation likelihoods and compensate numerical mismatches of the streams by employing a histogram equalization technique. Applied to a large audio-visual speech recognition task, the presented EQ yields lower sensitivity to the optimal choice of the stream weight and higher word accuracy at low SNRs.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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)*, Baltimore, MD, USA, Jul. 2003, pp. 605–608.

[6] Receveur, S.; Meyer, P.; Fingscheidt, T., "A Compact Formulation of Turbo Audio-Visual Speech Recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 1–5.

[7] 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.