On Dynamic Stream Weight Learning for Coupled-HMM-based Audio-visual Speech Recognition

Ahmed Hussen Abdelaziz, Lalla Amina Charaf, Steffen Zeiler*, Dorothea Kolossa

Institute of Communication Acoustics, 44801 Bochum, Germany

Email: {Ahmed.HussenAbdelAziz, Lalla.Charaf@rub.de, Steffen.Zeiler, Dorothea.Kolossa}@rub.de

Abstract

Visual speech features encoding lip movements are almost fully independent of acoustical environmental effects. Recently, they have therefore attracted significant attention for the purpose of robust automatic speech recognition, where they are typically deployed in conjunction with the conventional acoustical features. In order to optimally fuse audio and video features, the relative contribution of each modality to the recognition decision should be dynamically controlled, e.g. by so-called *stream weights*. Training stream weight estimators requires choosing a suitable feature-dependent or model-dependent reliability measure and an appropriate mapping function that maps this measure to the corresponding stream weight. In this paper, we compare different reliability measures and mapping functions for stream weight estimation, and we evaluate their performance in audio-visual speech recognition based on coupled HMMs for a range of adverse acoustical conditions.

Introduction

Using visual observations in conjunction with the conventional acoustical observations can increase the robustness of automatic speech recognition (ASR) systems in noisy environments. This is due to visual observations being almost independent of acoustical environmental effects such as additive and convolutive noise. To combine the video stream, i.e., recordings of the lip movements of the speaker's mouth, and the audio stream, different fusion models have been proposed. In this paper, we use a fusion model called the coupled hidden Markov model (CHMM) [1]. In CHMMs, the audio and video modalities are integrated at the state level. The overall score of a coupled state in a CHMM is computed via

$$b_{q_t=i}(O_t,\lambda_t) = \mathrm{p}\left(O_t^A|q_t^A=i^A\right)^{\lambda_t}\mathrm{p}\left(O_t^V|q_t^V=i^V\right)^{1-\lambda_t}. \tag{1}$$

The score of a coupled state $q_t = i = \{i^A, i^V\}$ in a CHMM given the audio-visual observation O_t at time frame t is evaluated in terms of the emission likelihoods $\mathrm{p}\left(O_t^s|q_t^s=i^s\right)\ s\in\{A,V\}$ of the audio and video states composing this coupled state. In (1), the so-called stream weight λ_t controls the contribution of each stream to the overall score of the coupled state according to the stream's reliability and its information content. The stream weights (SWs) should be adaptively adjusted in order to achieve a reliable performance gain.

In [2], a mapping function (MF) has been used to map one-dimensional acoustical reliability (confidence) measures (RMs) to frame-dependent stream weights λ_t. In this study, we compare three different acoustical reliability measures and five mapping functions to determine the optimal MF/RM combination for the stream weight estimation task. Two of the reliability measures are model-dependent, namely the entropy H and the dispersion D, and the third one, the signal-to-noise-ratio (SNR), is signal-dependent. The MFs are first- and second-order exponential functions, first- and second-order polynomial functions and sigmoidal functions.

In the following section, we present the definitions of all above reliability measures. We also describe the training algorithm used to estimate the mapping function parameters. The results obtained when applying the estimated stream weights to a CHMM-based audio-visual (AV) ASR system are shown in the last section and conclusions are drawn.

Stream weight estimation

The entropy H and the dispersion D are model-based reliability measures that can be estimated given the audio HMM as:

$$\mathrm{H}_t^A = \sum_{i=1}^{N^A}\mathrm{p}\left(q_t^A=i^A|O_t^A\right)\log\left(\mathrm{p}\left(q_t^A=i^A|O_t^A\right)\right)\ \text{and} \tag{2}$$

$$D_t^A = \frac{2}{L^A(L^A-1)}\sum_{k^A=1}^{L^A}\sum_{l^A=k^A+1}^{L^A}\frac{\mathrm{p}\left(q_t^A=k^A|O_t^A\right)}{\mathrm{p}\left(q_t^A=l^A|O_t^A\right)}. \tag{3}$$

Before computing the dispersion as in (3), the L^A largest posteriors $\mathrm{p}\left(q_t^A=k^A|O_t^A\right)$ should first be arranged in descending order. On the other hand, the posteriors of all N^A states of the audio HMM are used in (2) to estimate the entropy H. The frame-wise signal-to-noise-ratio can be estimated via:

$$\mathrm{SNR}_t = 10\log\left(\frac{S_t}{N_t}\right), \tag{4}$$

where S_t and N_t are the estimated signal and noise energies at time frame t, respectively. The signal and noise power estimates required to compute these energies can be obtained using algorithms like improved minima controlled recursive averaging (IMCRA) and a speech preprocessor like the Wiener filter.

*This work has been supported by the Ministry of Economic Affairs and Energy of the State of North Rhine-Westphalia, Grant IV.5-43-02/2-005-WFBO-009.

Each of the three RMs mentioned above are used as an argument of a one-dimensional function that maps them to a frame-dependent SW. The parameters of each mapping function are estimated in a supervised manner using least squared error (LSE) optimization. The training data, i.e., the input arguments and the target outputs, are estimated as follows: For each data set that contains speech signals recorded under same acoustical conditions, i.e. same noise type and level, one input/output tuple is found. The input is computed by averaging the chosen RM over all frames of the data set. Thus, all mentioned RMs have been extracted from the audio stream, which is sufficient here, since there is just one visual condition, comprising only high-quality video data. As a target SW output of each set, a global fixed stream weight is found via grid search, minimizing word error rate.

Experiments and Results

For evaluation, we have used the Grid audio-visual corpus. We have divided the signals into three sets: A training set containing 90% of the signals, and a development and a test set containing 5% each. The training set has been used to separately train the audio and video HMMs. The development set has mainly been used to train the parameters of the MFs. To test the proposed approach under different acoustical conditions, we have used eight additional noisy versions of the test and development set. The noisy signals have been created by adding babble and white noise signals to the clean signals at four SNR levels between 0dB and 15dB. The babble and white noise signals stem from the NOISEX-92 corpus and were chosen to represent both time-variant and stationary noise.

We have used the first 13 mel-frequency cepstral coefficients (MFCC) concatenated with their first and second temporal derivatives as the acoustical observations. The visual observations are 64-dimensional DCT coefficients encoding the appearance and shape of the speaker's mouth. The corresponding mouth region has been determined automatically by a Viola-Jones face and mouth detector. The dimensions of the acoustical and visual observations have finally been reduced to 31 using linear discriminant analysis (LDA).

The single-modality word HMMs are speaker-dependent, linear models. The number of states in each HMM is proportional to the number of phonemes contained in this word with a proportionality factor of 3 for audio HMMs and 1 for video HMMs. The output probability distributions of all emitting states are Gaussian mixture models with 3 mixture components for audio HMMs and 4 for video HMMs. The Java Audio-visual SPEech Recognizer (JASPER) has been used for training and recognition.

The results in Tables 1, 2, and 3 show that the first order exponential function with the dispersion as its input argument gives the best average performance. However, no single mapping function performs optimally for all reliability measures under all acoustical conditions. Therefore, more complex multiple-dimensional mapping functions will be investigated in future works.

Table 1: AVASR performance obtained using SWs mapped from the dispersions using five different MFs.

Noise Type	SNR [dB]	Poly. 1st	Poly. 2nd	exp. 1st	exp. 2nd	Sigm.
Babble	15	91.31	91.06	92.05	90.98	91.51
	10	86.71	86.44	87.60	86.56	86.69
	5	82.13	81.40	83.53	82.49	81.70
	0	80.88	80.68	79.98	81.97	79.88
White	15	91.54	91.22	92.52	90.94	91.81
	10	87.64	87.29	89.15	86.97	87.99
	5	85.07	84.58	86.28	84.59	85.32
	0	83.46	83.24	84.29	84.09	83.57
Clean	-	98.74	98.74	98.79	98.73	98.74
Avg.	-	87.50	87.18	88.24	87.48	87.47

Table 2: AVASR performance obtained using SWs mapped from the SNRs using five different MFs.

Noise Type	SNR [dB]	Poly. 1st	Poly. 2nd	exp. 1st	exp. 2nd	Sigm.
Babble	15	90.87	89.40	89.24	89.36	9082
	10	87.55	87.56	87.37	87.86	8783
	5	83.14	84.52	84.47	85.00	8426
	0	83.42	82.11	84.31	83.43	8430
White	15	88.27	86.68	86.54	86.87	8887
	10	82.52	78.60	78.42	78.63	8223
	5	80.68	80.19	80.22	81.60	8097
	0	81.73	83.68	83.61	84.02	8305
Clean	-	98.60	98.69	98.59	98.60	9863
Avg.	-	86.31	85.71	85.86	86.15	8677

Table 3: AVASR performance obtained using SWs mapped from the entropies using five different MFs.

Noise Type	SNR [dB]	Poly. 1st	Poly. 2nd	exp. 1st	exp. 2nd	Sigm.
Babble	15	90.38	90.28	91.59	90.51	90.64
	10	83.90	83.21	85.51	84.47	84.02
	5	77.18	73.91	79.05	78.29	76.80
	0	76.04	70.91	76.48	77.11	75.36
White	15	90.45	90.69	91.67	90.43	90.63
	10	85.31	84.76	87.13	85.69	85.26
	5	81.03	79.93	82.81	81.40	81.07
	0	79.73	77.33	81.22	80.47	79.68
Clean	-	98.88	98.91	98.85	98.83	98.88
Avg.	-	84.77	83.33	86.03	85.24	84.70

References

[1] A. 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. 1274–1288, 2002.

[2] V. Estellers, M. Gurban, and J.-P. Thiran, "On dynamic stream weighting for audio-visual speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1145–1157, 2012.