MAXIMUM ENTROPY AND MCE BASED HMM STREAM WEIGHT ESTIMATION FOR AUDIO-VISUAL ASR

Guillaume Gravier, Scott Axelrod, Gerasimos Potamianos, Chalapathy Neti

IBM T. J. Watson Research Center
PO Box 218, Yorktown Heights, NY 10598, USA
{ggravier, axelrod, gpotam, cneti}@us.ibm.com

ABSTRACT

In this paper, we propose a new fast and flexible algorithm based on the maximum entropy (MAXENT) criterion to estimate stream weights in a state-synchronous multi-stream HMM. The technique is compared to the minimum classification error (MCE) criterion and to a brute-force, grid-search optimization of the WER on both a small and a large vocabulary audio-visual continuous speech recognition task. When estimating global stream weights, the MAXENT approach gives comparable results to the grid-search and the MCE. Estimation of state dependent weights is also considered: We observe significant improvements in both the MAXENT and MCE criteria, which, however, do not result in significant WER gains.

1. INTRODUCTION

Audio-visual speech recognition (AVSR) has recently received a lot of interest, mainly because the visual modality is not affected by acoustic noise. Various models of audio-visual integration for speech recognition have been proposed, among which the multi-stream hidden Markov model (MSHMM) has been demonstrated to consistently improve recognition over audio-only ASR [1, 2, 3]. This model is based on the use of parallel HMMs to represent various streams of information. The parallel HMMs are considered independent and are re-synchronized at some pre-determined points to ensure the coherence of the overall process. In state synchronous MSHMMs, the stream HMMs share the same state sequence.

One of the key issues in multi-stream modeling is the combination of partial path stream scores when synchronizing the HMMs. A popular technique is linear combination of the stream log-likelihoods. However, this technique requires that the coefficients of the linear combination, known as the stream weights, be determined in some way. Stream weights can be fixed by hand to some values reflecting the relative confidence one has in a stream. Alternately, they can be estimated at training time or adapted at test time according to some measure of the reliability of each stream. Two popular techniques for doing so are brute force search and discriminative training, such as minimum classification error (MCE) estimation [6]. Here, we shall introduce an alternative approach based on a maximum entropy (MAXENT) criterion.

In most previous work, stream weights have been taken to be global, i.e. state independent (see however [4]). In this paper, we investigate the case of state dependent weights which many have thought would improve accuracy by reflecting the fact that the relative reliability of the different streams is dependent on the speech event. In the state dependent case, we only compare the MCE and MAXENT trained weights, since it it is impractical to perform a brute force search.

The paper is organized as follows. We first review the mathematical formulation of MSHMMs and then present the stream weight estimation algorithms used. Finally, experimental results are given and discussed in sections 4 and 5.

2. STATE SYNCHRONOUS MULTI-STREAM MODELING

2.1. Mathematical framework

Let us denote $y(t) = (y_a(t),\ y_v(t))$ the audio-visual feature vector at time t, where $y_a(t)$ and $y_v(t)$ are the audio and video feature vectors respectively. The conditional score of $y(t)$ given a state i is given by

$$\ln b_i(y(t)) = \sum_{s=a,v} w_{is} \ln b_{is}(y_s(t)) \ , \qquad (1)$$

where the stream pdf b_{is} can be any density function. Gaussian mixture densities are considered throughout the paper. It must be noted that (1) does not define a density, even under the constraint that $\sum_s w_{is} = 1$. To avoid confusion, we will refer to this function as a score. Finally, the state dependent stream weights can be tied in different ways, the two extreme cases being fully HMM state dependent weights and weights tied at the global level.

2.2. Parameter estimation

Two main parameter estimation strategies are possible with multi-stream models. The first strategy consists in estimating HMMs with identical topology independently for each stream. The HMMs are then *joined* into a multi-stream HMM. In the state-synchronous case, the stream density for each (possibly context dependent) state is taken from the corresponding state in the single-stream model. The transition probabilities are arbitrarily taken from one of the single stream models, the audio model in our case.

The second strategy consists in using the Baum-Welch algorithm to estimate the parameters of the multi-stream model jointly for each stream. The state occupation probabilities are calculated using the current multi-stream model and the parameters of each stream density b_{is} are re-estimated independently. Note that it is assumed that the stream weights w_{is} are known.

3. STREAM WEIGHT ESTIMATION

It can be easily shown that maximum likelihood (ML) estimation of the stream weights is not tractable unless some constraints stronger than the sum to one constraint are used [5]. However, such constraints are hardly justifiable. Therefore, stream weight estimation cannot be carried along with the ML estimation of the HMM parameters, so suitable algorithms must be devised. Discriminative criteria, such as the popular minimum classification error (MCE) criterion [6] or the maximum mutual information criterion [7], have been proposed in the literature. Another popular category of techniques consists in determining global weights from the signal to noise ratio (see, *e.g.* [8]).

We briefly recall the MCE algorithm and present a new technique based on the maximum entropy (MAXENT) criterion. This new method requires no parameter tuning and is very fast. It also provides a general framework within which various constraints on the weights can easily be implemented.

3.1. Minimum classification error

The MCE principle is to minimize an error function. Let $y^{(l)}$ be the audio-video data for the l'th training utterance of length T_l. Given a state sequence x, the average conditional log-likelihood per frame is given by

$$\mathcal{L}(y^{(l)}|x) = \frac{1}{T_l} \sum_{s=a,v} \sum_{t=1}^{T_l} w_{x(t),s} \ln b_{x(t),s}(y_s^{(l)}(t)) \ . \quad (2)$$

We define the utterance misrecognition measure $d(l)$ as the difference between the log-likelihood given the alignment $\hat{x}^{(l)}$ of the reference transcription $\hat{W}^{(l)}$ and a smoothed average of the log-likelihoods given the alignments of the competing sequences $W_n^{(l)}$. The competing hypotheses are taken as the N-best output of the recognizer. If $x_n^{(l)}$ is the state sequence corresponding to the n'th decoder hypothesis, the misrecognition measure is given by

$$d(l) = -\mathcal{L}(y^{(l)}|\hat{x}^{(l)}) + \ln \left(\frac{1}{u^l} \sum_{n=1}^{N} u_n^l \ \exp\left(\mathcal{L}(y^{(l)}|x_n^{(l)})\right) \right) \ . \quad (3)$$

The error function $e(l)$ is itself defined by

$$e(l) = \frac{1}{1 + \exp\left(-a(d(l) + b)\right)} \ , \quad (4)$$

where $a > 0$ and b are constants. In (3), $u_n^l = 1$ if the two word sequences $\hat{W}^{(l)}$ and $W_n^{(l)}$ are different and 0 otherwise, and $u^l = \sum_n u_n^l$.

Given the reference alignment $\hat{x}^{(l)}$, the N-best output $W_n^{(l)}$ and the corresponding alignments $x_n^{(l)}$, obtained with the current estimates $w_{is}^{(k)}$ of the weights, we shall estimate the weights using the following probabilistic gradient descent algorithm

$$w_{is}^{(k+1)} = w_{is}^{(k)} - \epsilon_k \frac{\partial e(l)}{\partial w_{is}} \ . \quad (5)$$

The steps ϵ_k are positive and slowly decrease towards 0. The crucial parameters of the MCE algorithm are the error function slope a, the initial step ϵ_1 and the decreasing speed of the ϵ_k sequence. A judicious choice for these parameters is necessary for convergence.

	training		held-out		test		
	#spk	#utt	#spk	#utt	#spk	#utt	#wrd
DIGIT	50	5,490	50	670	50	529	4,513
LVCSR	208	17,111	25	2,277	26	207	3,176

Table 1. Definition of the corpora on the DIGIT and LVCSR tasks.

3.2. Maximum entropy

The maximum entropy criterion [9, 10] dictates that the weights w_{is} should be chosen so as to maximize the posterior log-likelihood of a fixed reference alignment given the training audio-video vectors, i.e. we must maximize

$$\ln P(\hat{x}|y) = \sum_l \sum_{t=1}^{T_l} \ln b_{\hat{x}^{(l)}(t)}(y^{(l)}(t)) - Z_{l,t} \quad (6)$$

with respect to the stream weights, where

$$Z_{l,t} = \ln \sum_i b_i(y^{(l)}(t)) \ . \quad (7)$$

This is a convex optimization problem whose unique solution gives a conditional distribution $P(i|y(t))$ of maximum entropy, subject to certain constraints. Unlike the case of the MCE, the optimization can be performed fairly readily using any one of several generic optimization techniques. For this paper we used the publicly available quasi-newton optimization package L-BFGS-B (version 2.1) [11]. Virtually all of the computational time required is spent on the frame conditional likelihood computation. No decoding or "judicious" choice of parameters is necessary.

The above model is very flexible and may be generalized in many ways. As explained below, one generalization that we need to make use of for the LVCSR task is one in which the states are clustered. In this case, P no longer gives the probability of a state i, but rather of a cluster $\bar{\imath}$. In this case, we replace the function $\ln b$ in (1) by

$$\ln \bar{b}_{\bar{\imath}}(y(t)) = \sum_{s=a,v} w_{\bar{\imath}s} \max_i \ \ln b_{is}(y_s(t)) \ , \quad (8)$$

where the maximum is over all states i in class $\bar{\imath}$.

4. EXPERIMENTS

4.1. Evaluation tasks

Two tasks are considered in these experiments. The first one is a multi-speaker digit recognition task (DIGIT). The second one is a speaker independent large vocabulary dictation task (LVCSR). The experiments are carried out on artificially noisy audio data at various SNRs in a multi-condition training fashion, that is with models trained on the noisy data.

The DIGIT vocabulary comprises the digits from *one* to *nine* plus *zero* and *oh*. Utterances from 50 speakers were recorded and divided into three corpora as detailed in table 1. The LVCSR task is a 10k word speaker-independent continuous dictation task. As for the DIGIT task, three corpora are considered. In both tasks, speech babble noise was added to the original audio data at various SNRs, the video data remaining untouched.

The same set of features were used for both tasks. The audio stream features were obtained by training a linear discriminant

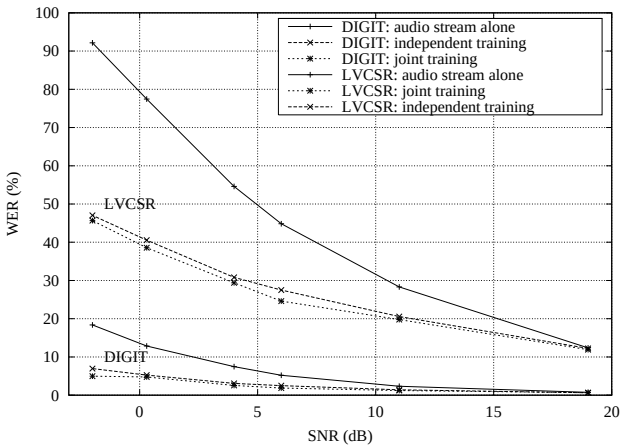

Fig. 1. WER on the DIGIT and LVCSR tasks using pure audio and audio-visual independently and jointly trained MSHMMs with global weights.

analysis (LDA) transforms of the cepstral coefficients, followed by a maximum likelihood linear transform (MLLT). The LDA and MLLT transform were trained for each noise condition. The video stream features were obtained by an LDA-MLLT transform of the pixels in a region of interest around the mouth as described in [2].

The audio-visual modeling is based on context dependent phone models. A set of 24 phones was used for the DIGIT task which correspond to 159 context-dependent states and 6k Gaussians. For the LVCSR task, 54 phones were considered with a total of about 2,808 states and 50k Gaussians.

4.2. Training of the audio-visual models

First, bootstrap audio and visual context-dependent HMMs were independently derived from generic audio models using a single pass retraining EM iteration.

Independently trained MSHMMs were obtained by first training single-modality HMMs from the bootstrap models by iterating the EM algorithm. Since the stream HMMs are derived from the same generic models, the audio and the video models share exactly the same topology and, in particular, the same context-dependent state clusterings. The stream HMMs were then joined to create the multi-stream models. Global stream weights were optimized for the DIGIT task by a grid-search procedure on the held-out set. The same global weights are used to join the LVCSR models at this stage. The reference alignments used in (6) were generated using these independently trained models.

Jointly trained MSHMMs were obtained by first joining the stream bootstrap models using the global weights determined for the independently trained models and by iterating twice the EM algorithm.

Figure 1 shows the word error rate on both tasks with the audio HMMs alone and with the independently and jointly trained MSHMMs. These results clearly demonstrate the advantage of audio-visual speech recognition in noisy conditions over the audio alone speech recognition. The advantage of jointly training the multi-stream model compared to an independent training of stream models is also clearly demonstrated.

4.3. Log-likelihood normalization

The log-likelihoods in each stream having very different ranges of values, they must be normalized before recombination when using state dependent weights. Indeed, if the video log-likelihoods are much higher than the audio ones, then any state having a high video weight compared to the other states would be privileged at decoding time. Therefore, the log-likelihoods should be normalized so that the average log-likelihood over all states is comparable in both streams. This is achieved by replacing the log-likelihood function $b_{is}(y_s(t))$ by

$$b'_{is}(y_s(t)) = \frac{b_{is}(y_s(t)) - \mu_s}{\sigma_s} \quad , \tag{9}$$

where μ_s and σ_s are the mean and standard deviation of $b_{is}(y_s(t))$. It was shown on both tasks that comparable results are obtained with raw and normalized log-likelihoods when global weights, optimized by a grid-search procedure, are used. This result means score normalization does not discard vital information and potentially enables the use of state dependent weights.

4.4. Weight estimation

As mentioned in section 4.2, the jointly trained MSHMMs were trained using global weights determined by a grid-search optimization using the independently trained MSHMMs. However, the weights used at training time may not be optimal for the new model. Also, it is interesting at this point to try to estimate state-dependent weights. The MCE and MAXENT algorithms described above were used to estimate both global weights and state-dependent weights on the held-out set for both tasks. A grid-search optimization of global weights was also carried out.

The experiments on the DIGIT task, partially illustrated in figure 2, demonstrated that the global weights trained with the MAXENT model performed comparably with the grid-search optimization. The MCE algorithm applied to the estimation of global weights turned out to be rather unstable and difficult to tune for convergence, probably due to the small score values obtained after normalization of the log-likelihood. Indeed, when non-normalized scores were used instead, the MCE algorithm converged and also gave results comparable to the ones obtained by grid-search on raw scores, though slightly worse. This point will be discussed further in the next section. Finally, the use of state dependent weights did not improve the performance of the system, whatever estimation algorithm was used, as can be seen from figure 2. However, it must be noted that the WER was not degraded by the use of state dependent weights, as was initially the case when raw log-likelihoods were used.

On the LVCSR task, global weights estimated either by MAXENT or by MCE gave results similar to the ones obtained with the grid-search optimization. No convergence problem of the MCE algorithm was observed on this task, even with normalized scores.

For MAXENT, some clustering of states was performed to avoid the problem of under representation of some states which were aligned to very few frames in the training sample set. First, the context-dependent states corresponding to the same context independent state were clustered together. Then states with fewer than 500 samples were clustered together, with states having between $k * 100$ and $(k + 1) * 100$ samples belonging to the same cluster.

The results obtained with state dependent weights are given in table 2 and compared to the results obtained with grid-search

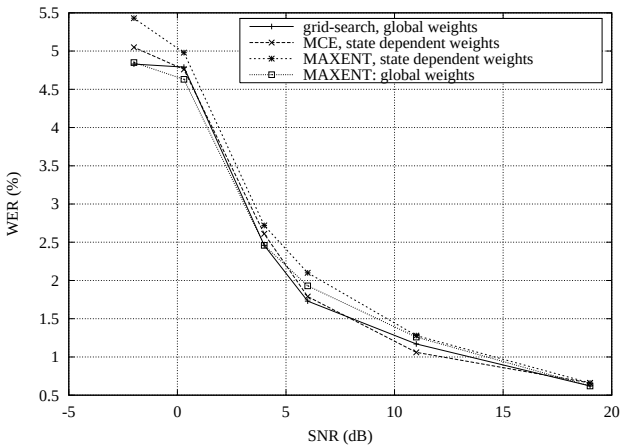

Fig. 2. WER on the DIGIT task for grid-search and MAXENT global weights, as well as MAXENT and MCE state dependent weights.

optimized global weights. As for the DIGIT task, no significant improvement was observed using MCE trained state-dependent weights. Surprisingly, the state dependent weights obtained by MAXENT did not perform well at the word error level in spite of a significant reduction of the frame classification errors observed on a subset of the LVCSR held-out set. For example, we observed a 28% relative frame error rate reduction at 6 dB.

SNR	20 dB	6 dB	0 dB
grid-search	12.3	24.9	38.8
MCE global	12.1	26.8	42.7
MAXENT global	11.7	24.9	39.1
MCE state	11.7	24.1	38.8
MAXENT state	13.7	32.4	46.0

Table 2. WER on the LVCSR task for global and state dependent weights.

5. DISCUSSION

One useful result that we have found is that the optimum global weights can be readily reproduced using the MAXENT technique, at a lower computational cost than grid-search or MCE. For the case of state dependent weights trained using MAXENT, both the MAXENT criteria function, as well as the frame error rate, showed significant improvement and yet, unfortunately, the WER degraded significantly. It is common knowledge that, in speech recognition, frame error rate improvements may not lead to WER improvements, but the fact that there was such a degradation in WER is surprising and interesting.

A problem with the MCE was that the optimal global weights found on the DIGIT task with non normalized scores did not match the optimal weights found by grid-search on the same held-out set. This result suggests that the smooth utterance error function (4) does not reflect accurately the word error rate, which is the measure we are actually trying to minimize. We verified this hypothesis by comparing the average utterance error function obtained with various values of the global audio stream weights to the WER obtained with the same weights. This comparison clearly demonstrated that the minima of the two functions do not match, in particular when the WER is very low.

As a final remark, we note that, although there has been a general sense in the community that the use of state dependent weights would be likely to improve word error rates, the results reported here suggest that this may not happen unless weight training is done using a criterion that is more tightly related to the WER. This motivates future experiments. For example, a better criterion for MAXENT training might be to use soft alignments of the training data to the true transcripts (i.e. posterior class probabilities from the Baum-Welch algorithm), rather than the hard alignments used in (6).

6. REFERENCES

[1] Stéphane Dupont and Juergen Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. on Multimedia*, vol. 2, no. 3, pp. 141–151, September 2000.

[2] Juergen Luettin, Gerasimos Potamianos, and Chalapathy Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, pp. 169–172, 2001.

[3] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Final Workshop Report, Center for Language and Speech Processing, 2000.

[4] Pierre Jourlin, "Word dependent acoustic-labial weights in hmm-based speech recognition," in *Proc. Audio-Visual Speech Processing*, pp. 65–68, 1997.

[5] Javier Hernando, "Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, pp. 1267–1270, 1997.

[6] Gerasimos Potamianos and Hans Peter Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, vol. 6, pp. 3733–3736, 1998.

[7] Yen-Lu Chow, "Maximum Mutual Information estimation of HMM parameters for continuous speech recognition using the N-best algorithm," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, pp. 701–704, 1990.

[8] Shigeki Okawa, Enrico Bocchieri, and Alexandros Potamianos, "Multi-band speech recognition in noisy environments," in *Proc. IEEE Intl. Conf. Acoust., Speech, Signal Processing*, pp. 641–644, 1998.

[9] E.T. Jaynes, "Notes on present status and future prospects," in *Maximum Entropy and Bayesian Methods*, Jr. W.T. Grandy and L.H. Schick, Eds. 1991.

[10] A. Berger, S. Della Pietra, and V. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, 1996.

[11] http://www-fp.mcs.anl.gov/otc/Tools/LBFGS-B.