Dynamic Modality Weighting for Multi-Stream HMMs in Audio-Visual Speech Recognition

Mihai Gurban, Jean-Philippe Thiran
École Polytechnique Fédérale de Lausanne
ELD 236, Station 11
1015 Lausanne, Switzerland

Thomas Drugman, Thierry Dutoit
Faculté Polytechnique de Mons
31 Boulevard Dolez
7000 Mons, Belgium

ABSTRACT

Merging decisions from different modalities is a crucial problem in Audio-Visual Speech Recognition. To solve this, state synchronous multi-stream HMMs have been proposed for their important advantage of incorporating stream reliability in their fusion scheme. This paper focuses on stream weight adaptation based on modality confidence estimators. We assume different and time-varying environment noise, as can be encountered in realistic applications, and, for this, adaptive methods are best-suited. Stream reliability is assessed directly through classifier outputs since they are not specific to either noise type or level. The influence of constraining the weights to sum to one is also discussed.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*signal processing, computer vision*

General Terms

Algorithms

Keywords

Audio-Visual Speech Recognition, Multimodal Fusion, Multi-stream HMM, Stream reliability

1. INTRODUCTION

Multimodal recognition tasks take advantage of the complementarity of different sources of information, in order to improve performance compared to monomodal classifiers. This fusion of information sources can be achieved in different ways. First, feature vectors can be merely concatenated and transformed, resulting in a multimodal feature vector. This is referred to as *feature fusion*. Alternatively, in *decision fusion*, the outputs of monomodal classifiers are merged to draw a final classification. Unlike the previous approach, decision fusion techniques have the ability to incorporate stream reliability to improve their robustness to adverse time-varying conditions. Furthermore, they can operate at different temporal resolutions, which allows asynchronicity to different degrees. This leads to the distinction between *early*, *intermediate* and *late integration*. In this paper we focus on a particular early integration technique using decision fusion via a weighted product rule: the state synchronous multi-stream Hidden Markov Model (MSHMM) [1].

The paper is structured as follows. Since our application is Audio-Visual Speech Recognition (AVSR), a brief overview of this topic, including our implementation details, is given in Section 2. Section 3 introduces the MSHMM principles. Section 4 investigates the stream weight estimation in an adaptive unsupervised way. More precisely the choice of modality confidence indicators relying on classifier outputs, as well as their mapping towards stream weights is discussed. Section 5 shows our audio-visual recognition results. Finally section 6 concludes and presents some guidelines for future work.

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.

2. OUR AVSR SYSTEM

AVSR is very well-suited as an application of multimodal fusion techniques. AVSR uses visual information derived from the video of the speaker to improve the audio speech recognition results, especially when the audio is corrupted by noise. This can be done because the audio and the video are complementary in this case, that is, the phonemes that are easily confused in the audio modality are more distinguishable in the video one, and vice-versa. On the other hand, the fusion methods used for AVSR are very general and can be implemented in any situation where several streams of information need to be combined.

For our experiments, we use sequences from the CUAVE audio-visual database [6]. They consist of 36 speakers repeating the 10 digits. Out of these 36 sequences, 30 are used for training and 6 for testing. The accuracy that we report is the number of correctly recognized words minus insertions, divided by the total number of test words, averaged over 6 runs with different train/test sets.

We use the HTK library [12] for the HMM implementation. Our word models have 8 states with one diagonal-covariance gaussian per state. The audio features used are 13 MFCCs, together with their first and second temporal derivatives, extracted 100 times per second. The temporal resolution of the video is increased through interpolation to reach 100 fps, since synchrony between the audio and the video streams is required by our integration method. The visual features are even-frequency discrete cosine transform

(DCT) coefficients of the mouth images, since they contain the information related to the symmetrical details of the image, as detailed in [9]. From them, the highest-energy 64 coefficients are selected, with their first and second temporal derivatives, and LDA is applied on them, to obtain a 40-dimensional feature vector.

3. THE MULTI-STREAM HMM

HMMs have the ability to model sequences of data, making them particularly well-suited for speech recognition. Multi-stream HMMs [1] have been proposed to generalize this framework for multimodal processing. They are similar to classical HMMs, with the particularity that each state contains not one, but several emission probability models, one for each stream, which are combined through a weighted product. In this way the emission likelihood b_j for state j and observation o_t at time t is expressed as:

$$b_j(o_t) = \prod_{s=1}^{S} [\sum_{m=1}^{M_s} c_{jsm} N(o_{st}; \mu_{jsm}, \Sigma_{jsm})]^{\lambda_{sjt}} \qquad (1)$$

where $N(o; \mu, \Sigma)$ denotes the value in o of a multivariate Gaussian with mean μ and covariance matrix Σ. For each stream s, M_s Gaussians are used in a mixture, each weighted by c_{jsm}. In Equation 1, the contribution of stream s is weighted by an exponent λ_{sjt} which in general can also depend on time t and state j. The latter dependency is not considered in this paper. In practice weights λ_{st} should be dynamically tied to stream reliability, such that, when environment conditions (e.g. SNR) change, they can be adjusted to emphasize the most reliable modality. In most previous work, stream exponents are constrained to sum to one, although this is not theoretically necessary. Since the multimodal emission likelihood in equation 1 is only a score, the weights are free of any restriction. Subsection 5.1 shows the gains that can be made by removing the constant sum constraint.

There are two possible ways to train MSHMMs. The first is separate training, where different models are built and trained for each modality. The two resulting HMMs are then merged into a MSHMM, with the gaussian mixtures from both models. However, there is no guarantee that the models will be trained on the same alignment of audio and video. The second method of training avoids this drawback, by using a joint multi-stream model from the beginning. Still, this requires an initial weight in the training phase. For our experiments we chose this method, with an initial weight of 0.5.

4. STREAM WEIGHT ESTIMATION

These last years a great interest has been devoted to the determination of stream weights for multimodal integration. Some methods employ a discriminative training relying on different criteria: minimizing the Classification Error or maximizing entropy in [4], an optimization based on likelihood value normalization in [11], or the use of a discriminative model combination in [3]. In [8] the authors build class specific models and anti-models in an unsupervised way. The stream weights ratio is then expressed as a non-linear function of intra- and inter-class distances. In all these methods, there is some training of the weights done on a held-out dataset. More intuitively weights have also been computed as a linear function of an estimated SNR [2].

Our approach consists in finding estimators of stream reliability and mapping them to stream weights. This is done dynamically, as our assumption is that noise can vary in time, and thus the recognition system should be able to adapt to changing conditions. Because of this we also avoid a supervised learning approach, as we assume that the choice of a held-out set for training has too big an influence on the results.

4.1 Stream reliability indicators

Our aim here is to define a coherent measure assessing a modality's reliability. In AVSR some audio indicators are extracted on the speech signal: the voicing index ([3]) or the SNR [2], and video corruption is generally neglected. Our approach estimates stream confidence directly from each classifier's outputs. It is assumed that if a clear peak emerges in the posteriors distribution, the stream is reliable. On the opposite, if classes have a flat posterior distribution, ambiguity is strong and the modality is unreliable. There are several ways of measuring the posterior probabilities' dispersion ([7],[5]). In [10] the maximum posterior probability is used as a reliability measure. We use the entropy, as it takes into account the whole probability distribution, not just the maximum.

However, using only the frame-level posteriors from time t might not be sufficient to assess the reliability of a stream, since speech classes are not local decisions, but temporal sequences. For the moment though, we are using the local estimates to obtain dynamically changing stream weights.

4.2 Mapping towards stream weight

Even after a suitable stream reliability measure has been found, the question of deriving stream weights from it remains. The weights and the reliability measure might not be on the same scale, and their relation might be non-linear. Some approaches use training on a held-out subset of the training dataset, but, as mentioned before, we try to avoid this, as we consider that the type and intensity of noise in testing conditions is uncertain. This means that having a held-out set for weights training that matches the target noise conditions is very unlikely.

The approach that we take is to find a mapping that satisfies a few basic conditions. In our case, as the posterior entropy is used as a reliability estimator, the relationship between the estimator and the weight is inverse, that is, when the entropy is low, the weight should be high, and vice-versa. We also impose for the moment that the sum of the weights should be 1.

Let the audio and video streams' entropies be H_a and H_v, and their associated weights λ_a and λ_v. The maximum value that the entropy can reach in our case is $H_{max} = log_2 83 \simeq 6.3$ since we have 83 classes. The mapping should ensure that when $H = 0$, $\lambda = 1$ and when $H = H_{max}$, $\lambda = 0$. Obviously, when $H_a = H_v$, $\lambda_a = \lambda_v = 0.5$.

There are several possible mappings that can be used. Two possibilities are presented below:

$$\lambda_a(t) = \frac{H_{max} - H_a(t)}{\sum_s H_{max} - H_s(t)} \qquad (2)$$

$$\lambda_a(t) = \frac{1/H_a(t)}{\sum_s 1/H_s(t)} \qquad (3)$$

We will refer to equations 2 and 3 as the "negative en-

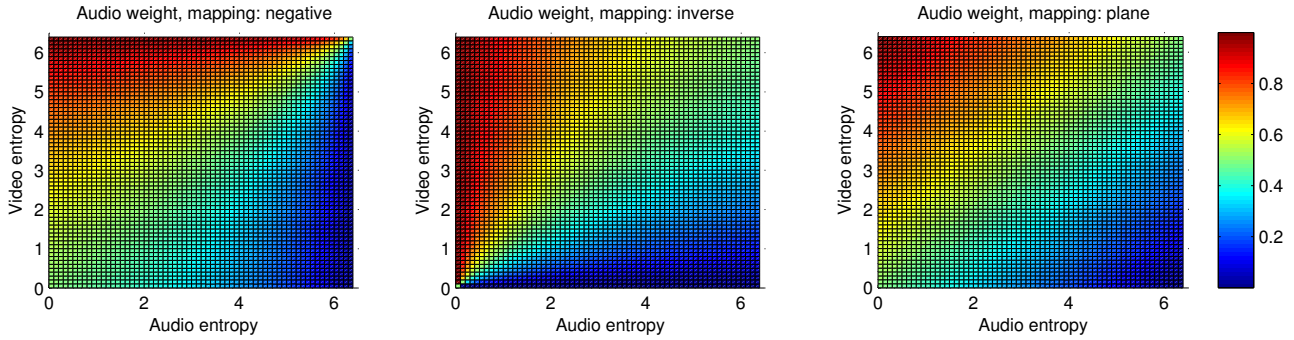

Figure 1: Three possible mappings from entropy to weight.

	sum=1	unconstrained	gain
clean	98.66	99.45	+0.79
10 dB	90.74	92.64	+1.90
-10 dB	57.18	61.93	+4.76

Table 1: Maximum performance with constrained and unconstrained weights, for three SNRs with babble noise.

tropy" and "inverse entropy" mappings. They have a common shortcoming: if one of the entropy values is close to an extreme (either zero or H_{max}), a variation in the other entropy's value will have no effect. This can be seen in figure 1. To avoid this problem, we derived a third mapping, which represents a plane in 3D space, as shown below:

$$\lambda_a(t) = \frac{H_v - H_a}{2H_{max}} + \frac{1}{2} \quad (4)$$

5. AV RECOGNITION RESULTS

In this section, we will present our results for AVSR, emphasizing the gain that can be obtained from multimodal fusion, especially in the presence of noise. We use two types of additive noise, white and babble. First, we show that allowing the weights to vary freely between 0 and 1, without constraining their sum, can lead to improved results. This is done with weights that are fixed for the whole length of the test sequence. Second, we present our results with dynamically adaptive weights based on stream entropy, as detailed in section 4.

5.1 Unconstrained static weights

Decoding in a HMM-based recognizer requires finding the most probable sequence of states given a series of observations, which is accomplished by searching the best path through a lattice of Markov states characterized by transition and emission costs. The final score of a path is a sum of weighted emission log-likelihoods and the logarithms of transition probabilities. Although the weights are only applied on the emissions, if their sum is allowed to be different than 1, they can change the balance of importance between the emission and the transition probabilities. A low value for this sum would mean that a higher importance is placed on the transitions, compared to the emissions, and vice-versa.

Figure 2 shows the impact of static (i.e constant during utterances) weights, on the Audio-Visual recognition rate, when the constraint on the sum is removed. Direction (D)

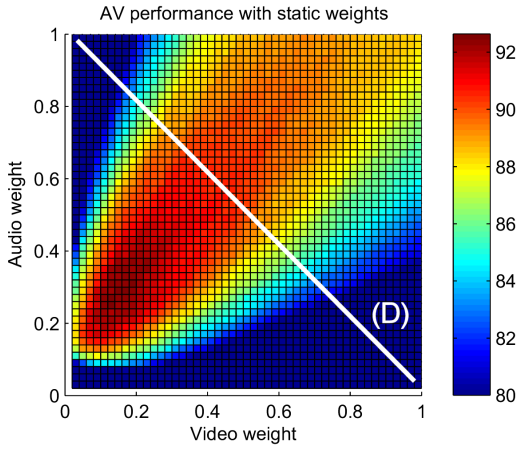

Figure 2: The influence of unconstrained stream weights on the recognition accuracy for $SNR = 10dB$ (babble noise).

represents the condition of unitary sum. Searching on this diagonal leads to a great improvement compared to monomodal results, however this may lead to missing the optimal performance point. In our example releasing this condition brings a gain of about 2% which is far from negligible. Table 1 shows that there are gains across all SNRs.

Even though the potential gain is obvious, finding the optimal scaling factor between emissions and transitions (i.e. the weight sum) is a difficult problem in the context of dynamically varying weights. Normally, emission likelihoods have a much higher order of magnitude (in log domain) than transition probabilities, and thus, they have a predominant influence on the decoding. However, in adverse conditions, they are less reliable and their importance should be reduced, as shown in figure 2. The fact that emission likelihood ranges also depend on the dimensionality of the feature vector makes this study even more complex, which is why it is the object of ongoing research. For now, the dynamic weights are constrained to sum to one.

5.2 Dynamic weights

Figure 3 shows the percentile accuracy obtained for audio-only and video-only recognition, as well as the three entropy mappings mentioned earlier. We also performed tests with another method from the literature, the maximum stream

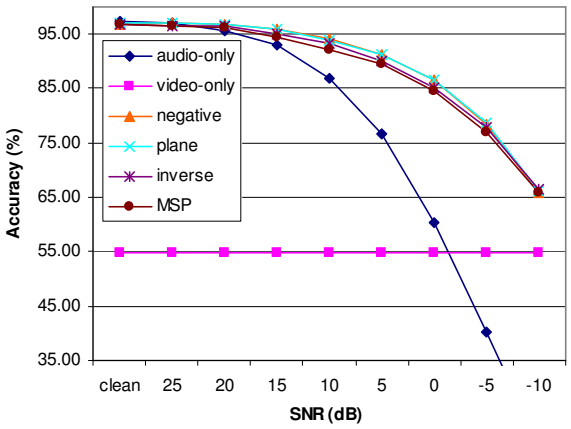

Figure 3: Word recognition results for both unimodal and multimodal recognizers, with different levels of white noise.

posterior (MSP) algorithm presented in [10], for comparison purposes. The MSP technique bases its choice of weights only on the maximum posterior probability at a given moment.

As can be seen, all multimodal methods outperform audio-only recognition by a margin that increases with the audio SNR, which was expected. The MSP method is the worst performing between the audio-visual ones, while the best entropy mapping seems to be the plane. The difference between the inverse entropy mappings and the other two shows that, all things being equal, the choice of the mapping function is also important.

6. CONCLUSION AND FUTURE WORK

We addressed the problem of stream weight estimation in the MS-HMM framework for AVSR. Our approach has two parts. First, we determined an estimator of stream reliability, in our case, the entropy of the state-level posteriors at each time instant. Secondly, we determined suitable mappings from the entropies to the stream weights, while avoiding training-based methods as they may not generalize well to unseen noise types and levels. Our technique was compared to another state-of-the-art method, leading to a slight improvement.

A second contribution of our paper is showing that the condition typically put on the weights sum is unnecessary, and there are gains that can be made by using this sum to balance the influence of the transition probabilities during decoding.

In the end, two questions still remain open and are the subject of ongoing research. First, the entropy of the instantaneous posterior probabilities for each frame may not be an adequate choice for this task. Indeed, we are trying to recognize patterns in temporal sequences, so a measure that takes into account a longer time frame might be more adequate. Secondly, allowing not only the weights but also their sum to vary dynamically might lead to better performance, if a suitable method of adaptation is found.

7. ACKNOWLEDGMENTS

Thomas Drugman is supported by the Belgian "Fonds National de la Recherche Scientifique" (FNRS). Mihai Gurban is supported by the Swiss National Science Foundation through the IM2 NCCR.

8. REFERENCES

[1] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. *Proc. International Conference on Spoken Language Processing*, pages 426Ú–429, 1996.

[2] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. In *IEEE Transactions on Multimedia*, volume 2, pages 141–151, Sept. 2000.

[3] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin. Weighting schemes for audio-visual fusion in speech recognition. In *ICASSP01, Salt Lake City, USA*, volume 1, pages 173–176, May 2001.

[4] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti. Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR. *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2002.

[5] H. Misra. *Multi-stream processing for noise robust speech recognition*. PhD dissertation, EPFL, Lausanne, May 2006.

[6] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. Moving-talker, speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus. *EURASIP JASP*, 2002(11):1189–1201, 2002.

[7] G. Potamianos and C. Neti. Stream confidence estimation for audio-visual speech recognition. In *ICSLP00, Beijing, China*, volume 3, pages 746–749, 2000.

[8] E. Sanchez-Soto, A. Potamianos, and K. Daoudi. Unsupervised stream weight computation using anti-models. In *ICASSP07, Hawaii, USA*, April 2007.

[9] P. Scanlon and G. Potamianos. Exploiting lower face symmetry in appearance-based automatic speechreading. *Proc. Works. Audio-Visual Speech Process. (AVSP)*, pages 79–84, 2005.

[10] R. Seymour, D. Stewart, and J. Ming. Audio-visual integration for robust speech recognition using maximum weighted stream posteriors. *Proc. INTERSPEECH*, 2007.

[11] S. Tamura, K. Iwano, and S. Furui. A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization. In *ICASSP05, Philadelphia, USA*, pages 468–472, March 2005.

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge, Entropic Ltd., 1999.