On Dynamic Stream Weighting for Audio-Visual Speech Recognition

Virginia Estellers, *Student Member, IEEE*, Mihai Gurban, and Jean-Philippe Thiran, *Senior Member, IEEE*

***Abstract*—The integration of audio and visual information improves speech recognition performance, specially in the presence of noise. In these circumstances it is necessary to introduce audio and visual weights to control the contribution of each modality to the recognition task. We present a method to set the value of the weights associated to each stream according to their reliability for speech recognition, allowing them to change with time and adapt to different noise and working conditions. Our dynamic weights are derived from several measures of the stream reliability, some specific to speech processing and others inherent to any classification task, and take into account the special role of silence detection in the definition of audio and visual weights. In this paper, we propose a new confidence measure, compare it to existing ones, and point out the importance of the correct detection of silence utterances in the definition of the weighting system. Experimental results support our main contribution: the inclusion of a voice activity detector in the weighting scheme improves speech recognition over different system architectures and confidence measures, leading to an increase in performance more relevant than any difference between the proposed confidence measures.**

***Index Terms*—Adaptive weighting, audio-visual speech recognition, multi-modal classification, multi-stream hidden Markov model (HMM), robust speech recognition, stream reliability, voice activity detection (VAD).**

I. Introduction

THE performance of automatic speech recognition (ASR) systems degrades heavily in the presence of noise, compromising their use in real world scenarios. In these circumstances, ASR systems can benefit from the use of other sources of information complementary to the audio signal and yet related to speech. Visual speech constitutes such a source of information. Mimicking human lipreading, visual ASR systems are designed to recognize speech from images and videos of the speaker's mouth. This fact gives rise to audio-visual automatic speech recognition (AV-ASR), combining the audio and visual modalities of speech to improve the performance of audio-only ASR, especially in presence of noise [1], [2]. In these situations we cannot trust the corrupted audio signal and must rely on the visual modality of speech to guide recognition, that is, give more importance to the visual than the audio cues when taking decisions about the speech classes. Consequently, the problem of weighting each of the modalities for speech classification naturally arises.

The weight assigned to each modality should be related to its reliability to classify speech. In a quiet environment with ideal audio and visual signals, higher weight should be given to the audio stream, reflecting the fact that the audio modality is more reliable than the video when it comes to recognize speech. When one of the modalities is degraded (due to background noise in the audio channel or an occlusion of the speaker's mouth) the importance assigned to it should decrease and reflect the confidence we have on that modality in such circumstances.

In general terms the problem can be formulated as the combination of different streams of information in a classification task and is therefore not limited to AV-ASR. Indeed, it has been introduced in biometric person identification [3] to include feature streams from different modalities and in multi-band speech recognition [4] to consider different processing techniques applied to the same audio signal.

In our work we focus on the integration of the audio and visual information for the recognition of speech. We propose a dynamic scheme where weights are derived from instantaneous measures of the stream reliability, some specific to speech processing and others inherent to any classification task. The use of fixed weighting schemes has already been addressed in AV-ASR literature [2], [5]–[19], but only a few works [20]–[24] focus on dynamic weights adapting the system to changing environmental conditions. Moreover, some of the results reported in literature for dynamic weights seem contradictory [20]–[22] and conclusions cannot be derived because different confidence measures have been tested with different AV-ASR architectures, recognition criteria, and databases. In this sense, the first contribution of the paper is a fair comparison of existing stream reliability measures in the estimation of the optimal weights. To this purpose we adopt the same form for the measure-to-weight mapping and optimization criteria and test the different confidence measures in both standard hidden Markov models and artificial neural network systems. The main contributions of the paper are, however, a new confidence measure inspired by the Viterbi algorithm and the introduction of a voice activity detector (VAD) in the weighting scheme, taking into account the special role of the silence class in the definition of stream weights in AV-ASR systems. In fact, our experiments show that the improvement associated to the introduction of a VAD in the definition of stream weights is more relevant than any difference

Manuscript received November, 2010; revised August, 2011; accepted October 06, 2011. Date of publication October 17, 2011; date of current version February 10, 2012. This work was supported by the Swiss SNF under Grant 200021-130152. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chung-Hsien Wu.

The authors are with the Signal Processing Laboratory LTS5, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. (email: virginia.estellers@epfl.ch; mihai.gurban@epfl.ch; jp.thiran@epfl.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2172427

of performance between the proposed stream confidence measures.

The rest of the paper is organized as follows. In Section II, we explain how the audio and visual integration takes place in ASR systems, review different stream weighting techniques proposed in literature and justify the necessity of dynamic weights adapting the system to changing environmental conditions. In the context of dynamic weights, in Section III we present existing techniques to estimate their correct value as a function of the stream reliability and propose a new confidence measure. In Section IV, we explain how the proposed confidence measures are mapped to stream weights and justify the use of different weighting strategies for the speech and silence intervals. In Section V, we report experiments with a reference database, comparing the performance and limitations of the different weighting techniques. Finally, conclusions are drawn in Section VI.

II. Multi-Modal Fusion for AV-ASR

In this section, we present the state-of-the-art for audio-visual fusion and stream weighting in ASR. We do not attempt to review the literature of ASR and refer the reader to [25]–[29] for a more complete overview of speech recognition models. We simply justify the weighting model adopted in speech recognition and explain how it is included in the audio-visual models. The last part of the section focuses on the weights associated to each stream as parameters of the model and reviews the existing techniques for their estimation.

A. Weighted Multi-Stream Classifiers

In statistical classification it is common to assume that features of different streams are independent of each other. In this case, the statistical models factorize the joint probability distribution into single-stream distributions and reduce the complexity of the system. In the case of the audio-visual speech, perceptive studies showed that humans treat the streams as class conditionally independent [30], [31], that is, audio and visual features o_A, o_V are independent given that the speech class $q = q_i$. Under this hypothesis, speech recognition can be improved by introducing stream weights λ_A, λ_V and probability combination rules [32]. The model commonly used is a weighted geometrical combination of the audio and visual likelihoods, i.e.,

$$p(o_A, o_V|q = q_i) = p(o_A|q = q_i)^{\lambda_A}\, p(o_V|q = q_i)^{\lambda_V}. \quad (1)$$

This model controls the importance of each modality in the classification task with its associated weight[33] and includes the hypothesis that audio and visual modalities are class conditionally independent (equivalent to setting unitary weights $\lambda_A = \lambda_V = 1$).

Introducing the previous weighting schemes into the statistical models used in ASR leads to the definition of multi-stream hidden Markov models (HMMs) [25]. In single-stream HMMs, a discrete state variable $q(t)$ evolves through time as a first-order Markov process and controls the observed features $o(t)$ by defining a statistical model for the emission likelihoods $p(o(t)|q(t) = q_i)$[1]. An HMM therefore factorizes the problem into the estimation of transition probabilities between states, which encode the temporal evolution of speech, and emission likelihoods associated to each state.

In multi-stream HMMs, only the emission likelihoods are affected by the inclusion of different streams. The likelihoods are now computed independently for each stream and combined at a certain level, which depends on the integration technique. In early integration the streams are assumed to be state synchronous and the likelihoods are combined at state level as indicated by (1). Late integration, in its turn, combines the likelihoods at utterance level, while in intermediate integration the combination takes place at intermediate points of the utterance. The weighting scheme, nonetheless, remains the same and early or intermediate integration are generally adopted as leading to better results and finer control of the stream integration [21]. A common restriction is that the weights λ_A, λ_V sum up to one, which comes from the factorization of the probability associated to any HMM state sequence into transition probabilities and state emission likelihoods. In multi-stream HMMs, $\lambda_A + \lambda_V = 1$ assures that the ratio between the logarithm of the emission likelihoods and transition probabilities is kept the same as in single-stream HMMs, which are the units of measure used in the expectation maximization or Viterbi algorithm for recognition.

In speech, the transition probabilities are chosen to force the HMM to evolve from left to right while either generative or discriminative strategies are used to estimate the probability distributions of the observed features. In generative systems a separate probabilistic model, usually a Gaussian mixture model (GMM) [27], is assumed for $p(o(t)|q(t) = q_i)$ and the corresponding parameters of the model are separately estimated for each class q_i. On the other hand, discriminative models use a single artificial neural network (ANN) or support vector machine (SVM) to assign a class probability distribution to the observed data $P(q(t) = q_i|o(t))$ and are thus designed to discriminate between classes, not to generate class models. In this sense, training ANNs or SVMs to classify speech from different classes is more complex than estimating independently the GMMs for each class, but leads to models computationally simpler at testing stage than a large collection of GMMs. We will see that the use of GMM or ANN also affects the definition of stream reliability measures based on the performance of the classifier.

B. Weight Estimation Criteria

If we assume that the weights are fixed parameters of our models, we can estimate their optimal value with training or held-out data. In this case, the trained weights will only be relevant for the particular environmental conditions in which that data was acquired.

Ideally, we want to choose the weights that minimize the final word error rate (WER) of our classifier, which is the natural measure of performance in ASR. However, the WER is not a

[1]We use P to indicate the estimated probability value from a discrete distribution and p for the value taken by a probability density distribution of a continuous variable

smooth function of the training data, as its computation involves finding the most likely path between all possible state sequences and penalizing different types of errors (insertions, deletions and substitutions). Therefore, using the minimum WER as optimization criterion leads to simple grid-search methods choosing the weights with minimum WER in a training dataset, as reported in [7], [8].

The WER is a global measure of the performance of the system. It gives a score for each utterance, but it does not reflect the temporal evolution of the error within the speech sequence or how the weights affect the likelihood of the speech models used for classification. To overcome that issue, some authors have proposed different smooth measures of the system's performance [5]–[8], [24], allowing the use of standard iterative techniques and optimization criteria on the training dataset. Those techniques usually minimize the frame error rate and maximize the discrimination between the different hypothesis of the classifiers. For instance, in [5], [7], HMMs are used to find the N most likely state alignments for the training data and their associated audio and visual likelihoods. The weights are then chosen to maximize the discrimination between the incorrect state alignments (from the N most-likely alignments associated to each sequence in the training data) and the correct one in terms of their joint audio-visual likelihoods. It is not clear, however, how those measures of the system's performance relate to the final WER. Indeed, in [7], the authors point out that the minimum WER of their training dataset and the optimum of their proposed smooth function are not obtained for the same value of weights.

Other methods do not involve a training procedure. The weights are not chosen to optimize the WER or any function of the system's performance on some training data, but are set at testing to adapt the system to the working conditions based on the data itself. In [9], the authors use previous theoretical results [10] to estimate the optimal stream weights as inversely proportional to the single stream misclassification error. To that purpose, they build class specific models and anti-models and use them in a small amount of unlabeled data to compute inter and intra-class distances for each stream, from which they estimate their classification error and the corresponding optimal weights. Another criterion is proposed in [11], where the weights are chosen to maximize the dispersion of the test emission likelihoods and lead to a more discriminative classification, even thought they might cause a wrong recognition. An extension of this algorithm is based on output likelihood normalization [12], where class-dependent weights are computed as the ratio between the average class-likelihoods over a time period. Note that here the weights become dynamic, as they are defined to normalize the class likelihoods at each time instant.

Dynamic stream weights, however, are usually introduced to adapt the system to changing environmental conditions due, for instance, to the temporary presence of noise in one stream. In this case, we cannot estimate the weights as fixed parameters of the system, but we have to make them evolve as a function of the estimated noise on each channel and the reliability of that stream for classification. For each noise level or estimated stream reliability, the weights can be chosen to optimize different measures of the performance of the system: recognition of isolated words [13]–[15], WER of continuous speech [2], [22], [23] or frame classification error [20], [21]. The weights can then be adjusted based on the estimated signal-to-noise ratio (SNR), as in [2], [13]–[17] or the voicing index in the case of speech recognition [18], [22], [34]. Other weighting methods applied to recognition tasks are modality-independent, as they are determined by the classifier's confidence [19]–[21] and can be used indifferently with the audio, video or any other stream. Only a few of the previous works [20]–[24] applied dynamic weights to AV-ASR, that its allowing the weights to change at frame level and adapt dynamically to the different noise conditions within a sequence. It is also interesting to note that a few works [23], [24], [35], [36] combined audio and visual estimates of the stream reliability to define audio and visual weights.

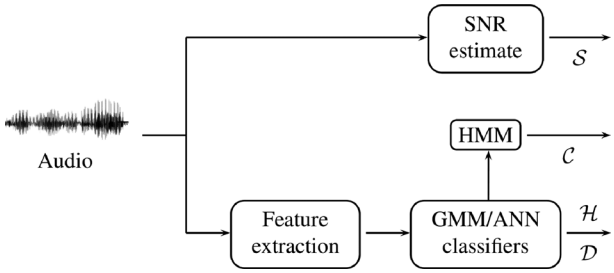

Fig. 1. Reliability of the audio stream can be estimated directly from the signal by measuring the noise present on the channel (with an estimate of the SNR $\mathcal{S}$ from Section III-A), analyzing the estimated posterior probability distributions of the classifiers (with measures like the entropy $\mathcal{H}$ or dispersion $\mathcal{D}$ from Section III-B) or taking also into account the time evolution of the HMM speech models (with the proposed measured $\mathcal{C}$ from Section III-C).

In our work, we focus on the use of dynamic weights in different AV-ASR systems (HMM-ANN and HMM-GMM) when the audio stream is subject to noise. We present existing confidence measures, propose a new and computationally simpler one and study how they map to the weights leading to minimum WER in a noisy training dataset. The minimum WER criterion is chosen because it is the final measure used to evaluate the system's performance and, as we have already said, it is not clear how other criteria relate to it. The questions that naturally arise are, for instance, deciding how to weight the audio and video streams during the silence periods inherent to speech, how quickly to adapt those weights in relation to the variations of the noise present on the stream and how the final WER is affected by the use of dynamic weights in a controlled noisy environment. The current paper shows the advantages, limitations and restrictions necessary to apply dynamic weights with changing levels and types of noise and points out the importance of silence recognition in the weighting scheme.

III. Measuring Stream Confidence

Two main strategies exist to estimate the reliability of the audio stream during fusion, either estimating a measure of the noise present on the channel directly from the audio signal or analyzing the estimated posterior probability distributions of the classifiers. Fig. 1 shows a block diagram of the proposed schemes.

Based on the speech signal itself we obtain an estimate $\mathcal{S}$ of the SNR present on the audio channel by means of VAD and simple power estimates. To measure the classifiers confidence

we use the dispersion $\mathcal{D}$ or the entropy $\mathcal{H}$ of the class posterior probabilities and HMM emission likelihoods. We show how each measure is implemented and suits HMM-ANN or HMM-GMM systems and propose a new measure $\mathcal{C}$ common and suitable for both architectures.

A. SNR of the Audio Signal

Measuring the SNR in ASR requires estimating the power of speech (signal of interest) and the power of the noise present on the audio signal. Due to the bursting nature of speech, we can obtain estimates of the power of the noise during the silence intervals inherent to any utterance and derive from it an estimate of the power of the speech signal. The estimated SNR, denoted as $\mathcal{S}$, is then computed as the ratio of the speech and noise power estimates. To this purpose, we must first detect the silence and speech intervals with a VAD. At each time instant, we compute the power of the audio signal and assign it to speech or non-speech. If the sample is associated to non-speech, we use it to update an estimate of the noise power. Otherwise the sample is detected as speech and, assuming noise and speech to be independent, the power of speech is estimated subtracting the previous estimated power of noise from the power of the audio signal. Note that estimating the power of noise during silence intervals defines an artificial low SNR for the silence intervals, when actually no speech is present and the SNR is ill-defined. This has a non-negligible effect on the weighting strategy, as the experiments will show.

As the aim of our work is not the study of VAD, we choose the VAD technique best suited to our system. We justify our choice as follows. Most VAD systems have a training stage where speech and non-speech models are built, usually assuming different Gaussian probability distributions for the speech and noise samples of some features related to ASR. On testing, a hypothesis test is used to estimate the likelihood of each sample belonging to speech or non-speech [37] and the results of the classification are afterwards smoothed in time to avoid short-time jumps and assure a minimum duration of words and silences. The results of this instantaneous classification can be smoothed with an HMM and, in fact, we can directly use an audio-only HMM-GMM (which is part already of our AV-ASR system) to segment the signal into speech and non-speech intervals. First, the audio-only HMM-GMMs of our system estimate the likelihood of each sample belonging to a phoneme or silence class. Afterwards, the standard Viterbi decoder, together with the estimated HMM transition probabilities, vocabulary and grammar are used to recognize speech, that is, to partition the utterance into a sequence of words and silence intervals as required from a VAD. Note that using the estimated SNR as a confidence measure is not particular to the audio channel, but its computation by means of a VAD is limited to speech signals. Moreover, it assumes a non-speech nature of noise and is therefore not designed to cope with babble noise. The use of audio-only HMMs, instead of other energy or GMM-based VAD techniques, provides a more robust detection of speech/silence intervals when the audio signal is subject to babble noise originated from a different vocabulary than the trained HMM. Similarly, the inclusion of an out-of-vocabulary detector (not used in our system), could improve the performance of the VAD and, consequently, the quality of the SNR estimate.

B. Confidence Measures of the Classifier

Generally it is advantageous to use stream confidence measures based on the classifier itself, as they convey information about the reliability of the data for the classification task, can be applied to any stream and are not specific to audio or speech signals. In ASR systems, the distribution of the posterior class probabilities or data likelihoods are the most common confidence measures derived from the classifiers. These measures assume that if the classifier assigns a very high probability to a certain class while the rest present low probabilities, then the sample being tested fits correctly one of the trained models and the classification can be considered reliable. Conversely, when all classes have similar probabilities or emission likelihoods, the sample does not seem to distinctively fit any particular class and we assume it is corrupted by noise or due to an unreliable stream. We want to point out that it is a reasonable assumption for speech classifiers trained with generative criteria (GMM usually), but its validity with ANN systems trained to discriminate between classes is less clear and, in fact, has proved false in experiments with very noisy data. Nevertheless, two measures have been proposed in ASR literature to capture this information: the entropy and dispersion of the posterior class probabilities or data likelihood of single-stream HMM classifiers.

In HMM systems, the dispersion of the emission log-likelihoods was first proposed in [13] to measure the difference on the probability scores of the N most likely states. Formally, if $\{q^1(t)\dots q^N(t)\}$ are the sorted N most likely states for the audio stream at time t, then the log-likelihood dispersion associated to the stream is

$$\mathcal{D}(t)=\frac{2}{N(N-1)}\sum_{m=1}^{N}\sum_{n=m+1}^{N}\log\frac{p(o(t)|q(t)=q^m(t))}{p(o(t)|q(t)=q^n(t))}. \tag{2}$$

The dispersion measures how distinguishable the different classes are in terms of emission likelihoods. High values of dispersion are associated to a small level of confusion in the classifier and a reliable stream, whereas a low dispersion is encountered when all the likelihoods take similar values and the classes are highly confusing. An equivalent dispersion measure has been defined with class posterior distributions instead of emission likelihoods [22].

The entropy of the state posteriors has also been used both in HMM-GMM [20], [21] and in HMM-ANN systems for multiband [38] and Audio-Visual ASR [22]. It is defined as

$$\mathcal{H}(t)=-\sum_{i=1}^{L}P(q(t)=q_i|o(t))\log P(q(t)=q_i|o(t)) \tag{3}$$

where $\{q_1\dots q_L\}$ are now all possible HMM states. It is important to note that in HMM-GMM systems the estimation of state posteriors requires the use of the Bayes rule and estimates of the prior class probabilities from the training dataset. Contrarily to the dispersion, the entropy reaches its maximum value for equiprobable classes and has low values when the sample seems to specially fit one of the classes.

Even though both entropy and dispersion are based on measuring the peakiness of the probability distributions, it is not clear which one is better suited for the task of stream reliability estimation. In [20] and [21] HMM-GMM systems obtained better performance for the dispersion than the entropy measures, while the contrary was observed with HMM-ANNs [22]. These apparently contradictory results are due to the effect of estimating class prior probabilities when computing the entropy in HMM-GMM systems and to the different training strategies (generative training of GMMs compared to a discriminative training of ANNs). In terms of implementation, the entropy can be directly computed in the ANN models, which have class posterior probabilities as output. In the HMM-GMMs, however, the Bayes rule must be first applied for the computation of the entropies, while the dispersion can be directly computed from the emission log-likelihoods. In that sense, entropy seems more adequate for the ANN than the GMM architecture, while dispersion of posteriors or log-likelihoods suits both models equally. Nevertheless, computing the dispersion requires sorting the instantaneous likelihoods or probabilities and is computationally more expensive than the entropy. In our work we use both dispersion and entropy to measure the reliability of ANN and GMM-based HMM systems and propose a new measure suited to both architectures and computationally simpler.

C. Proposed Confidence Measure of the HMM Classifier

We observe that both GMM and ANN systems share the same HMM structure to control the time evolution of the speech, but that only the GMM and ANN outputs were used in the definition of entropy and dispersion. We propose a new measure of the classifiers confidence not based on the values taken by the GMM or ANN's emission likelihoods, but on the transition probabilities of their common HMM structure. The proposed measure is inspired by the Viterbi decoder, where the transition probabilities between neighboring sequence states are combined with their emission likelihoods to find the most likely sequence of states, naturally including the left-to-right property of speech HMMs and vocabulary restrictions. Our measure takes into account both the data likelihood and the time evolution constraints inherent to speech and exploits single-stream classifiers in terms of GMM/ANN models and HMM transition probabilities. These two terms can also be understood as a measure of data fidelity (emission likelihoods or class posterior probabilities associated to each sample) and a regularity constraint (transition probabilities associated to the most likely state of consecutive samples for each stream).

During recognition and for each stream we keep track of the most likely state in the single-stream ANN or GMM at each time instant $q^{ML}(t)$ (different from the most-likely state in the multi-stream HMM, whose computation requires the definition of weights) and accumulate the value of the transition probability between the previous and the instantaneous most likely state for the stream in $\mathcal{C}(t)$

$$\mathcal{C}(t) = \mathcal{C}(t-1) + p\left(q(t) = q^{ML}(t) \,|\, q(t-1) = q^{ML}(t-1)\right). \tag{4}$$

In practice, we do not keep track of the whole history of $q^{ML}(t)$, but define a limited memory to adapt the system to changing conditions. The transition accumulator $\mathcal{C}(t)$ is then implemented as a moving average of the transition probabilities between the instantaneous most-likely state for each stream. Note that a similar procedure of tracking the most-likely state and updating the log-likelihood of the path with the associated transition probability is done on the Viterbi decoder when recognition is performed with the single-stream HMM. In our transition accumulator, we do not keep track of the GMM/ANN emission likelihoods and simply use them to select $q^{ML}(t)$ and update the accumulator with the corresponding transition probability. The proposed measure is then easier to implement and suits both GMM and ANN architectures. Compared to entropy or dispersion, it does not require sorting or additional functions of the emission likelihoods; a max search and a single addition are enough to update of the transition accumulator.

The reliability associated to the stream increases with the value of the transition accumulator. If there is noise in the audio stream, its most likely state at each time instant will jump between states not matching the time evolution of the trained models and the associated transition probabilities will remain close to zero. We have experimentally observed that, in presence of noise, $q^{ML}(t)$ mostly jumps between states corresponding to impossible transitions for the system (from the first state of phoneme A to the second state of phoneme B, transitions not allowed by left-to-right HMMs and vocabulary restrictions) and thus the accumulator is updated with a transition probability equal to zero. This fact also justifies our choice of directly adding transition probabilities instead of their logarithms (compared to Viterbi), which does not provide a physical meaning to our measure, but results in a more stable confidence measure. Indeed, the proposed $\mathcal{C}$ avoids the instabilities and overflows of a transition counter accumulating the logarithm of transition probabilities close to zero. In terms of a regularity constraint, it corresponds to choosing a penalty function which allows punctual misfit of the data to the models (transition probabilities 0) instead of introducing a large penalty for them.

Compared to the entropy or dispersion, the proposed accumulator takes also into account the temporal evolution of speech and vocabulary restrictions (sequences not allowed in the vocabulary have transition probabilities equal to zero). Entropy or dispersion only consider the emission likelihoods of the GMM or ANN, that is, how a sample instantaneously fits the observation models but not how the sequence of features fit the time evolution of speech. In our proposed method, the observation models are used to choose the most likely state at each time instant (data fidelity), while the time evolution of speech is taken into account by the transition probabilities (regularity term). Moreover, the transition accumulator suits HMM-GMM and HMM-ANN architectures and is computationally cheaper to compute than the entropy or dispersion.

IV. From Stream Reliabilities to Weights

In this section we study how to map the previous stream reliability measures to the optimal weights, that is the weights giving minimum WER in a noisy environment. As expected, the different confidence measures reflect the changes in the level of noise present on the signal, but they are also affected by the presence of silence intervals inherent to speech. In this sense,

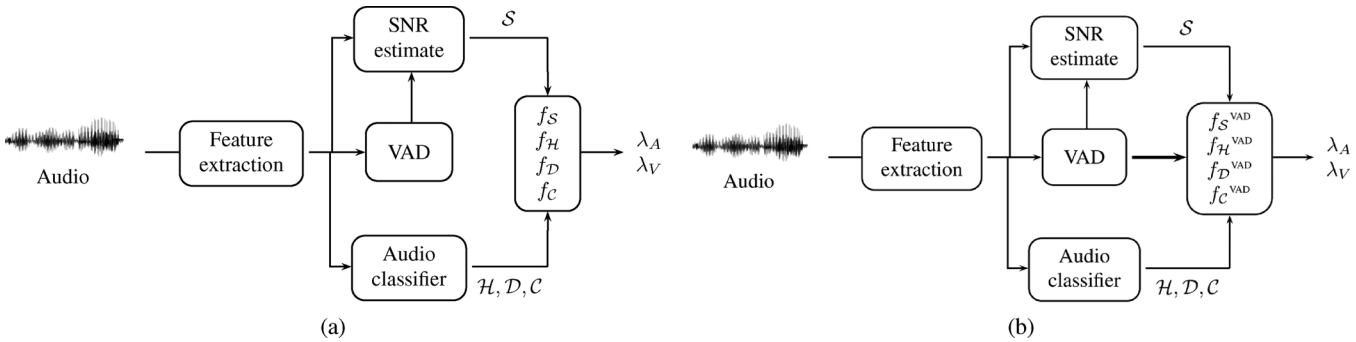

Fig. 2. Mapping of the stream reliability measure to optimal weights in the proposed system. For instance, for the transition accumulator $\mathcal{C}$, (a) uses a common mapping $f_{\mathcal{C}}$ for all the classes, while (b) proposes a different mapping for the detection of silence and speech $f_{\mathcal{C}}^{\text{VAD}}$. (a) Single mapping of stream reliabilities to weights. (b) Double mapping of stream reliabilities to weights.

an important contribution of this paper is the introduction of a VAD in the weighting system, showing that performance of all the reliability measures can be improved taking into account the special role of silence detection in AV-ASR. Fig. 2 shows the structure of the proposed weighting systems.

We start by mapping the different stream reliability measures to the stream weights leading to minimum WER in a noisy environment. To this purpose, we train the weighting system with an evaluation dataset subject to different known noise conditions. In the evaluation set we have artificially added different kinds and levels of noise to the audio stream. For each SNR level $n \in \mathcal{N}$ and type of noise, we do a grid search for λ_A^n, λ_V^n obtaining minimum WER on the evaluation data and satisfying $\lambda_A + \lambda_V = 1$. For the same dataset, we compute the mean value of the different stream reliability measures at each SNR level $\overline{\mathcal{S}}^n$, $\overline{\mathcal{D}}^n$, $\overline{\mathcal{H}}^n$, $\overline{\mathcal{C}}^n$, and define the mappings $f_{\mathcal{S}}$, $f_{\mathcal{D}}$, $f_{\mathcal{H}}$, $f_{\mathcal{C}}$ as continuous functions minimizing the mean square error (MSE) over all noise levels. For instance, for the transition accumulator we write

$$f_{\mathcal{C}} = \arg\min_{f_{\mathcal{C}}} \sum_{n \in \mathcal{N}} \left| \lambda_A^n - f_{\mathcal{C}} \left(\overline{\mathcal{C}}^n \right) \right|^2. \tag{5}$$

In order to fairly compare all the stream reliability measures, we define mapping functions of the same complexity for the different measures. From the values taken by those measures and the optimal weights in the evaluation data, we chose a weighted sum of exponentials for the mapping function $f(x) = A_1\, e^{B_1 x} + A_2\, e^{B_2 x}$, whose parameters A_1, B_1, A_2, B_2 are estimated iteratively with a region-trust method [39]. We chose that particular form for the mapping based on some preliminary results with the evaluation data. In reduced experiments with the evaluation data, the speech recognition performance obtained with the sum of exponentials, sigmoid[22], or piecewise linear functions[20] were comparable, but the values of the MSE resulting in (5) for the different measures were more similar to each other with the sum of exponentials than with the other functions. Using a sum of exponentials for the mapping function is therefore suitable to fairly study the differences between the stream reliability measures in the final speech recognition system, as it provides a similar performance in the estimation of the parameters of the continuous mappings $f_{\mathcal{S}}$, $f_{\mathcal{D}}$, $f_{\mathcal{H}}$, $f_{\mathcal{C}}$ from (5) for all confidence measures.

The average measures behaved as expected, with the dispersion and transition accumulator decreasing as the SNR level decreases and the entropy increasing for noisy data. It is to note, however, that the computation of the entropy in the GMM case is considerably sensitive to the estimation of the state prior probabilities and that the correct performance of that method requires a fine estimation of these probabilities. The best results are obtained using the time durations of phonemes in the training data to compute the state priors, while assuming equal class probabilities for all the states leads to a considerably poorer performance. Actually, as the distribution of phonemes in the training, evaluation and testing data is the same, the estimated priors match the testing ones. However, it is not generally the case in real scenarios and it is more advisable to use the transition accumulator than the entropy in HMM-GMM systems, which performs similarly, is simpler to compute and does not require the estimation of prior probabilities.

It is important to note that we try to learn a mapping to be applied dynamically and yet we estimate it by experiments with fixed weights. For the evaluation dataset we can justify it because the SNR is carefully kept fixed through all the sentences by artificially adding noise to the clean audio sequences. The value of the stream reliability indicator, however, varies within the sequence and we need to average it to define the mapping. In a real system, actually, the reliability measures and stream weights change instantaneously and the mapping learned from fixed weights might be incorrect. Smoothing the stream reliabilities through the testing sequences can give a similar behaviour to the one seen on training, but it does not ensure that the weights are instantaneously the best ones. In fact, it is necessary to study how each stream reliability measure evolves through an evaluation sequence with a fixed SNR level. If the confidence measure takes a relatively constant value throughout the sequence, then the mapping defined with the evaluation dataset between the mean value of this measure and the fixed optimal weights can be used. Otherwise, if the variations of the reliability measure on a fixed SNR sequence are comparable with the variations between different SNR levels, then the mapping cannot be directly used without a large smoothing of the confidence measure on the testing sequences, which hinders a quick adaptation to changing noise conditions.

Analyzing the evolution of the different reliability measures on the evaluation dataset, we state that the estimated SNR $\mathcal{S}$ re-

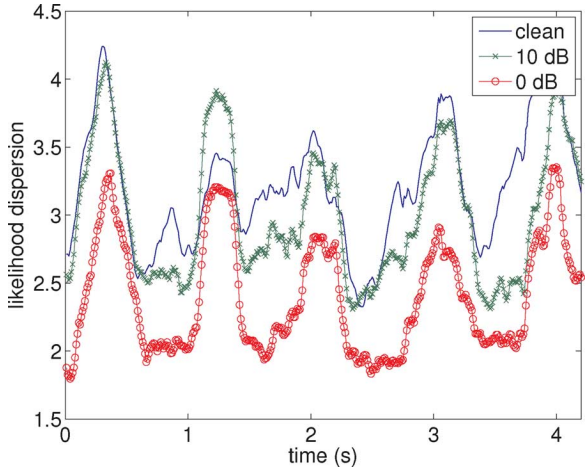

Fig. 3. Time evolution of the GMM likelihood dispersion for the same evaluation sequence and different levels of white audio noise artificially added. The likelihood increases at 1.2, 2, 2.9, and 3.8 seconds due to word utterances compared to its value during inter-word silences.

quires considerable smoothing, mainly due to the estimation of SNR during the silence periods between words. In those periods, the SNR ratio is small as there is no speech signal present. For the same sequence, the variations of $\mathcal{S}$ between the speech and non-speech intervals are higher than between the different SNR levels. Nevertheless, as $\overline{\mathcal{S}}^n \in \mathcal{N}$ evolves coherently through the different SNR levels, that issue can be solved assuring that the smoothing applied on testing always includes speech and silence intervals. The measures based on the classifiers confidence $\mathcal{D}$, $\mathcal{H}$, $\mathcal{C}$ show also different mean values for the silence and speech utterances, as shown in Fig. 3. The variations here are not caused by the SNR estimation procedure and indicate that the classifiers show different behavior for the speech and silence intervals.

This analysis suggests that the silence intervals inherent to speech might play an important role in the definition of proper stream weights. Indeed, in [22] the authors point out that for very noisy environments, training the weights with the minimum WER criterion leads to choosing the modality better suited for the detection of the silences existent between the words in the utterances. To avoid that kind of behavior, we developed a second strategy shown in Fig. 2(b). Using the same VAD used for the SNR estimation, we first assign each sample to speech or silence and then set the weights accordingly with different speech and silence mapping functions. To this purpose, we define two mappings from each reliability measure to the optimal stream weights, one for the recognition of silences ($f_{\mathcal{S}}^s$, $f_{\mathcal{D}}^s$, $f_{\mathcal{H}}^s$, and $f_{\mathcal{C}}^s$) and another for the recognition of speech ($f_{\mathcal{S}}^v$, $f_{\mathcal{D}}^v$, $f_{\mathcal{H}}^v$, and $f_{\mathcal{C}}^v$). To train this combined weighting system, we split the evaluation dataset into speech and silence examples based on the available labels, we concatenate them into continuous speech and silence utterances and learn the corresponding mappings as previously explained. Note that we learn the mappings from continuous speech recognition experiments with sequences containing only continuous speech or silence examples. Learning those mappings from isolated word recognition tests would define a weight threshold leading to correct or incorrect recognition of each word instead of minimizing the WER, which is the performance criterion used in continuous speech recognition. Now the stream reliability measures, specially the SNR estimator, are not influenced by the presence of silence intervals and are more stable within the evaluation sequences. Moreover, considering only speech or silence utterances to define the optimal stream weights, we obtain a mapping better suited to classify speech while the decision about silence or speech intervals is taken with a VAD designed to that purpose. We refer to these mapping strategies as $f_{\mathcal{S}}^{\mathrm{VAD}}$, $f_{\mathcal{D}}^{\mathrm{VAD}}$, $f_{\mathcal{H}}^{\mathrm{VAD}}$, and $f_{\mathcal{C}}^{\mathrm{VAD}}$ and note that they correspond to defining two different mappings from the stream reliabilities to the optimal weights: one for the detection of silence and another for the rest of speech classes. In case of the transition accumulator, for instance, we have

$$f_{\mathcal{C}}^{\mathrm{VAD}}(x) = \begin{cases} f_{\mathcal{C}}^s(x), & \text{if VAD classifies } x \text{ as silence} \\ f_{\mathcal{C}}^v(x), & \text{otherwise.} \end{cases} \tag{6}$$

On the other hand, the proposed $f_{\mathcal{S}}$, $f_{\mathcal{D}}$, $f_{\mathcal{H}}$, and $f_{\mathcal{C}}$ use a common mapping for all the classes and assume that the same value of weights is correct to differentiate silences from speech and to differentiate the speech classes between each other.

At testing stage, both single and double mapping weighting schemes first smooth the different reliability measures, then use the estimated mappings to compute the corresponding weights and use them in the multi-stream HMM classifier. Compared to the training stage, the confidence measures are not averaged per sequence but only smoothed in time. A moving average is used for smoothing the obtained $\mathcal{H}$, $\mathcal{D}$, and $\mathcal{S}$, while it is intrinsic to the definition of $\mathcal{C}$ as a moving transition accumulator. The window size adopted for the smoothing has been chosen based on preliminary experiments with the evaluation dataset, where a 20-ms window obtained good results across the different reliability measures. The other main difference between testing and training affects only the double-mapping schemes $f_{\mathcal{S}}^{\mathrm{VAD}}$, $f_{\mathcal{D}}^{\mathrm{VAD}}$, $f_{\mathcal{H}}^{\mathrm{VAD}}$, and $f_{\mathcal{C}}^{\mathrm{VAD}}$, where a VAD is used to classify samples as speech or silence. During training this decision is taken based on the labels of the evaluation data, while at testing a VAD is used for that purpose, which introduces a possible source of error.

The proposed double mapping strategy must not be confused with a class-dependent weighting scheme, in which the different multi-stream HMMs would have different stream weights depending on its class. In our system, at each time instant, the same audio and visual weights are used in all HMMs for audio-visual classification. For instance, if the VAD has detected a silence, the double mapping function for the transition accumulator will use $f_{\mathcal{C}}^s$ to estimate the optimal audio and visual weights, which will be used in all the multi-stream HMMs (irrespective of whether the HMM is associated to the silence or to a phoneme class).

V. Experiments

We perform continuous speech recognition experiments on the CUAVE database [40]. We use the static portion of the individuals section of the database, consisting of 36 static subjects repeating the digits five times in front of a camera. We do speaker independent experiments with 6-fold cross validation, using 30 speakers for training, 3 for evaluation, and 3 for testing. The results are given in terms of speaker independent WER and

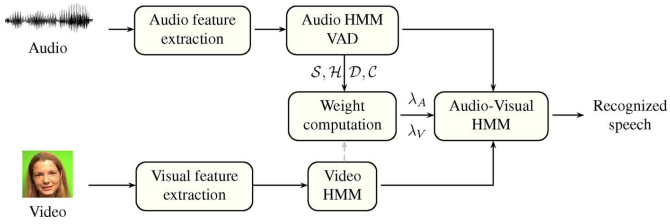

Fig. 4. Structure of the AV-ASR system. Audio and visual features are first extracted from the signals, which are passed to single-stream audio and visual classifiers to compute different stream reliability measures. These confidence measures are used to determine the associated audio and visual weights in the multi-stream classifier, where speech recognition takes place. In our experiments, the visual stream is assumed ideal and, consequently, no confidence measures are computed for that stream.

the statistical significance of the results is evaluated in a paired manner comparing the different confidence measures.

The block diagram of the proposed AV-ASR system is presented in Fig. 4 and the details about the different feature extraction, classification and weighting blocks is given in the following paragraphs.

A. Feature Streams

Normalized MFCC are extracted as audio features, with their first and second temporal derivatives. Thirteen MFCC features are computed with a 30-ms window, with 20-ms overlap, leading to an audio rate of 100 feature vectors per second. We train any HMM parameters on clean audio data and artificially add white and babble noise from the NOISEX database [41] on testing. Different levels of noise are added in order to show how our dynamic weighting algorithm performs across a large range of SNRs, from clean to -10 dB. Adding noise to the recorded signal instead of adding it during the recordings does not take into account the changes in articulation speakers produced when background noise is present [42] and therefore generates somehow nonrealistic scenarios. On the other hand, it enables to test exactly the same utterances in different noise conditions, facilitates the recordings of the data and complete control of the noise conditions in the evaluation set.

The visual features are selected discrete cosine transform (DCT) coefficients from a region of interest, which consists of a square of 128×128 pixels centred on the mouth, normalized for size, centred, and rotated. The DCT coefficients are the 13 most important ones taken in a zig-zag order, as in the MPEG/JPEG standard, together with their first and second temporal derivatives with their means removed. More details about this visual feature extraction system can be found in [43]. The temporal resolution of the visual features is then increased through interpolation to reach the audio rate, since synchronous audio and visual feature streams are required by the classifiers. No noise is added to the visual features as AV-ASR with non-ideal visual conditions requires the development of new visual feature extraction methods before audio-visual integration can be successfully studied [44], [45].

B. Speech Classifiers and VAD

We use the HTK library [46] for the HMM-GMM implementation of three-state left-to-right phoneme models. Each state has three Gaussians for the audio and one for the visual stream, all with diagonal covariance matrices. We start by training separately audio and visual HMM-GMM models, we then build the multi-stream models and jointly re-estimate their parameters setting the audio and visual weights to one during training.[2]

In the ANN case, the emission likelihoods are replaced by posterior phoneme probabilities estimated with a multi-layer perceptron. The audio and visual neural networks are implemented as feed-forward ANN with two neural layers and 10 000 neurons. One feature vector is feed to the ANN each time with sigmoid functions used in the input layer. The ANN has an output node for each class, with softmax functions used to provide an estimate of the class posterior probabilities associated to the input sample. The values of the transition probabilities from the HMM-GMM case are kept for the HMM-ANN system, as they correspond to a time model of the duration of phonemes learned from the same training data.

Recognition is based on phonemes, which are concatenated to form words and sentences by means of a dictionary and grammar. In our case, as the testing corresponds to sentences containing sequences of numbers, no grammar is used and the dictionary includes only the phonetic transcription of the English digits.

Designing the VAD we must compromise between having voice detected as noise or noise detected as voice (between false positive and false negative). In our case, the VAD must be able to detect speech under several types and levels of background noise and we design it to be fail-safe, that is, to detect speech when the decision is in doubt and lower the chance of missing speech segments. As already explained, we use the audio-only HMM systems to classify features as speech or silence. Single-stream HMM are also used to estimate the entropy, dispersion and transition accumulator reliability measures of the audio stream. The obtained confidence measures are used to compute the weights of the multi-stream HMM systems, taking also into account the decision of the VAD in the case of double-map weighting schemes $f_{\mathcal{S}}^{\mathrm{VAD}}$, $f_{\mathcal{D}}^{\mathrm{VAD}}$, $f_{\mathcal{H}}^{\mathrm{VAD}}$, and $f_{\mathcal{C}}^{\mathrm{VAD}}$.

At this point, we must highlight that the final speech recognition is not performed in two passes, a first pass guided by the VAD to detect between silence and speech and a second stage where the different phoneme models and vocabulary are considered. We do not adopt such a strategy because in that case the visual stream is not taken into account for the detection of silence, which has proved a useful strategy [48], [49]. In this paper the VAD is only used in the weighting strategies $f_{\mathcal{S}}^{\mathrm{VAD}}$, $f_{\mathcal{D}}^{\mathrm{VAD}}$, $f_{\mathcal{H}}^{\mathrm{VAD}}$, and $f_{\mathcal{C}}^{\mathrm{VAD}}$ to determine the use of a mapping designed for speech or silence intervals.

C. Evaluation of the Results

In our experiments we compare different weighting strategies learned and tested on the same data and the results, therefore, reflect differences between the weighting strategies rather than differences in the test datasets. In this case, the statistical significance of the results cannot be evaluated by means of confidence intervals associated to the performance of each method independently, but requires the comparison of the different methods in a

[2]Experiments have shown that final performance of the system is dominated by the value of the weights at testing and not during training [47]

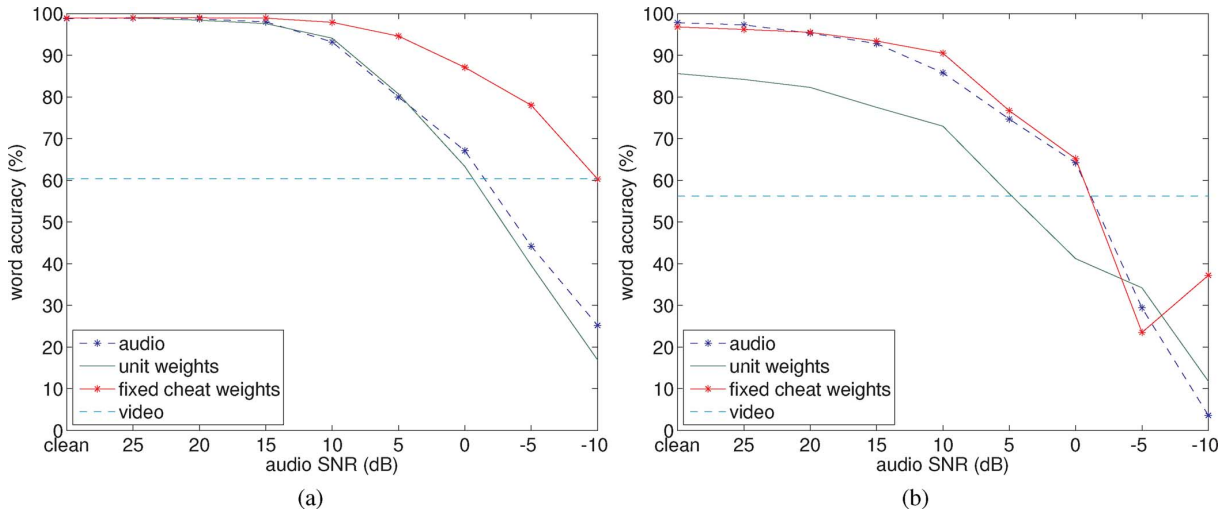

Fig. 5. Performance of single and multi-stream HMM systems for different SNR levels of white noise. The weighting strategy is significantly useful below 25 and 20 dB of audio SNR for the HMM-GMM and ANN case. (a) HMM-GMM base-line systems. (b) HMM-ANN base-line systems.

one-to-one basis for the same sentences, speakers and train/test datasets.

In speech recognition, a small modification to a system will alter the recognition results in a few sentences or speakers only. Intuitively we would acknowledge a 10% probability of reducing the errors if the number of errors drops on 10% of the sentences while the others remain unchanged. On the other hand, an overall improvement of the word error rate should be considered random if 50% of the sentences improved while 50% degraded. In this work, we use the "probability of error reduction" p_e presented in [50] to assess the differences in performance of the proposed weighting schemes. We refer the reader to the original paper [50] for a detailed description of p_e and present here only the main ideas. Intuitively, we measure the probability of error reduction p_e between two systems A and B counting the number of independent testing samples (sentences in our experiments) that favor system A over B while leaving the rest of the samples unchanged. Formally, however, the computation of p_e requires the estimation of the probability distribution associated to the paired comparison of the systems. To that purpose, we bootstrap the WER obtained by the different weighting methods for independent samples, count the number of samples favoring each system and perform a paired hypothesis test to obtain p_e. Bootstraping allows us to estimate the unknown distributions associated to the WER and the paired comparison of the systems and obtain an estimate of p_e which does not depend on the number of sentences used in each comparison and is defined in the same terms used to evaluate speech recognition systems. On the following, only the values of p_e relevant to assess if one method significantly outperforms another are given.

D. Experimental Results

The aim of this work is not to compare HMM-ANN and HMM-GMM architectures, so the results of the stream reliability measures will be compared for each of the systems separately. We include results for three extra baseline systems that we use to analyze the improvement obtained with a weighting strategy.

TABLE I
PERFORMANCE OF THE VAD FOR DIFFERENT TYPES AND LEVELS OF ACOUSTIC BACKGROUND NOISE. RESULTS ARE GIVEN IN TERMS OF TRUE POSITIVE RATE OR SENSITIVITY (TPR), FALSE POSITIVE RATE (FPR), AND ACCURACY (ACC)

white noise	clean	25dB	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	-10 dB
TPR	97.3	94.4	91.7	88.1	82.9	76.7	70.2	58.0	36.8
FPR	4.6	2.9	2.4	1.7	1.3	0.9	0.6	0.6	1.1
acc	96.3	95.8	94.8	93.4	91.2	88.4	85.5	79.7	69.7
babble noise	clean	25dB	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	-10 dB
TPR	97.1	94.8	92.7	89.5	82.5	73.9	61.2	45.9	36.1
FPR	4.6	3.6	3.5	5.0	8.4	15.3	23.5	28.8	29.9
acc	96.2	95.6	94.6	92.3	87.1	79.4	69.1	59.3	54.53

As mentioned previously, the use of a VAD for the computation of some reliability measures and weights introduces a source of error when the voice activity recognition fails. The performance of the VAD, shown in Table I, should also be taken into account in the analysis of the results. We observe tha the VAD works reasonably well for all levels of white noise and down to 5–0 dB for babble noise, when performance drops under 70%.

The first baseline system is an audio-only ASR system, showing the gain obtained by inclusion of the visual modality, and the other two are AV-ASR systems with fixed unit and "cheat" weights, that is weights assuming class conditional independence of audio and visual streams and the weights obtaining minimum WER for each SNR with the test dataset (with a common mapping for all the classes learned in the testing data, not on the evaluation set). Fixed unit weights corresponds to a system where no stream weighting is used and audio and visual features are just considered class conditionally independent. Comparing to such a system shows the improvement obtained by a weighting strategy under different noise circumstances. In Fig. 5(a) and (b), we see that a weighting strategy is significantly useful below 25 and 20 dB of audio SNR for the HMM-GMM and ANN case. The probability of error reduction p_e ranges from 0.7 to 1.0 for the different SNR levels and it defines the range of noise levels, where the comparison of the different weighting strategies is relevant. In that case, it is also important to note the performance of 60.4% obtained with a HMM-GMM visual-only system and 56.2%

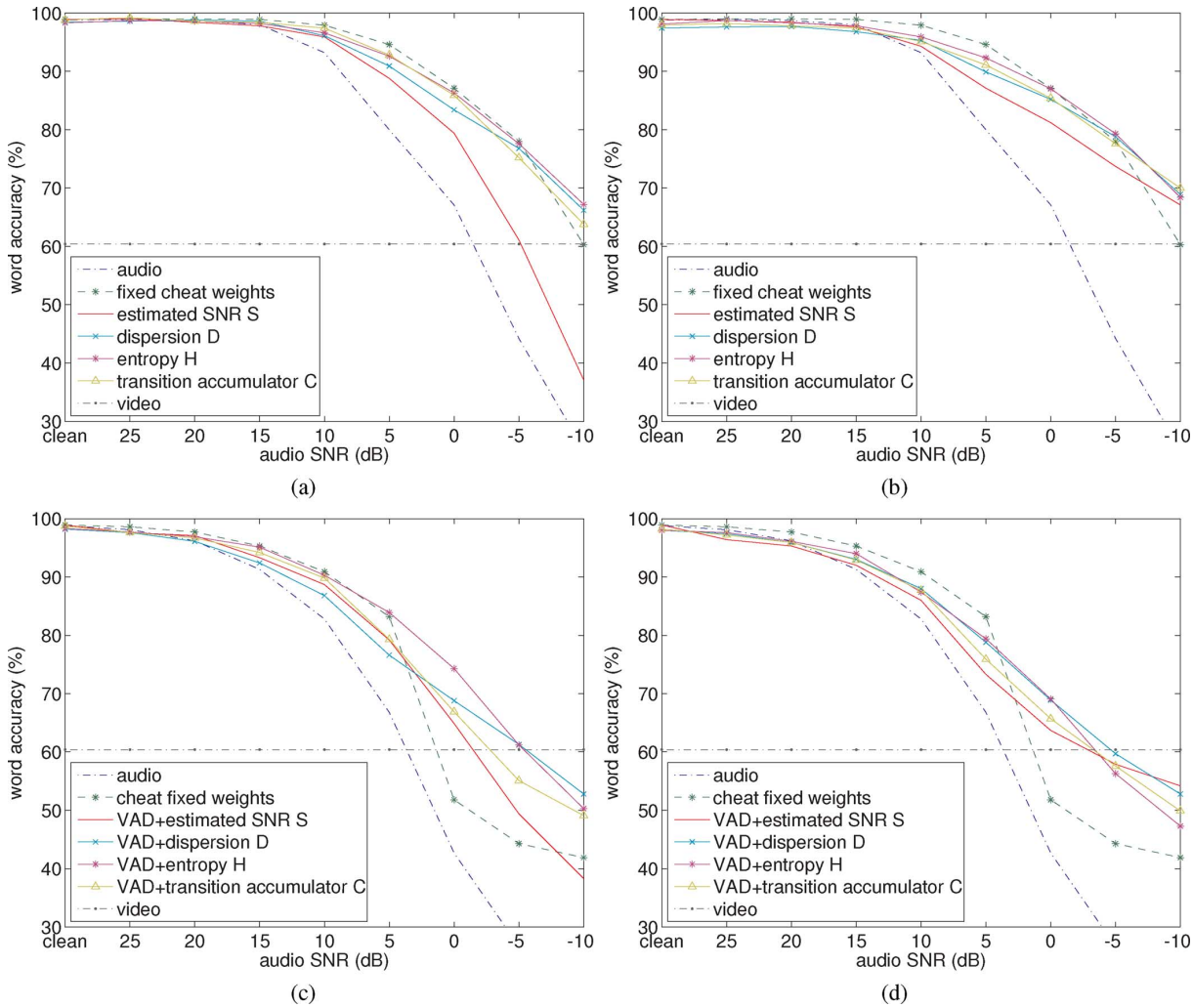

Fig. 6. Performance of cheat fixed weights HMM-GMM system and dynamic weights obtained with one and two mappings. Comparing to the fixed "cheat" system shows how far we are from the best behavior under the assumption that the weights depend only on the SNR. When a dynamic system outperforms the fixed one, it is due to the fact that the silence/speech class should also be taken into account for the weight definition. Comparing (a) with (b) and (c) with (d) we observe the improvement obtained by the inclusion of the VAD in the GMM system for white and babble noise. (a) One mapping, audio with white noise. (b) Two mappings, audio with white noise. (c) One mapping, audio with babble noise. (d) Two mappings, audio with babble noise.

for the HMM-ANN, which specially justifies the inclusion of the visual modality under 0–5 dB of SNR for the different systems and the study of AV-ASR in these circumstances. We observe that for audio SNRs, specially with babble noise, the audio-visual accuracy drops below visual-only accuracy even with the fixed "cheat" weights. It is explained by the fact that the weighting strategy is only able to compensate for uncorrelated errors in the audio and visual stream, but when both streams incur in the same kind of errors the audio-visual system also fails and its performance can be worse than the best of single-stream systems. In the case on low audio SNR, the audio system incurs in errors due to the false detection of silences, which is also a class easily confused in the visual domain and requires audio information to be detected. In such circumstances, the audio-visual system is not able to detect silences with "cheat" weights, which assume the same fixed weight can be used for the detection of silences than speech utterances. On other experiments we will show that the double mapping scheme partially overcomes this effect.

Comparing to the fixed "cheat" system shows how far we are from the best behavior under the assumption that the weights only depend on the SNR of the stream. In that sense, when a dynamic system outperforms the fixed one, it is due to the fact that the silence/speech class should also be taken into account for the weight definition, which is not the case with the "cheat" weights. Such is the case for the HMM-ANN system under 10 dB of SNR, see Fig. 7(b) and (d), with a probability of error reduction p_e over 0.9 for all the noise kinds and levels. The same behavior is observed under -5 dB and 5 dB SNR for the HMM-GMM case subject to white and babble noise, see Fig. 6(b) and (d). In these cases the optimal fixed weights choose the modality better suited for the silence detection, while the confidence measures including the VAD define different mappings for the silence and speech intervals. In this case, the performance of the AV-ASR is limited by the performance of the VAD, as shown by the experimental results.

Comparing the two mapping strategies ($f_{\mathcal{S}}^{\mathrm{VAD}}$ against $f_{\mathcal{S}}$, $f_{\mathcal{H}}^{\mathrm{VAD}}$ against $f_{\mathcal{H}}$, etc.), see Figs. 6 and 7, we observe that a considerable improvement is obtained when different mappings are

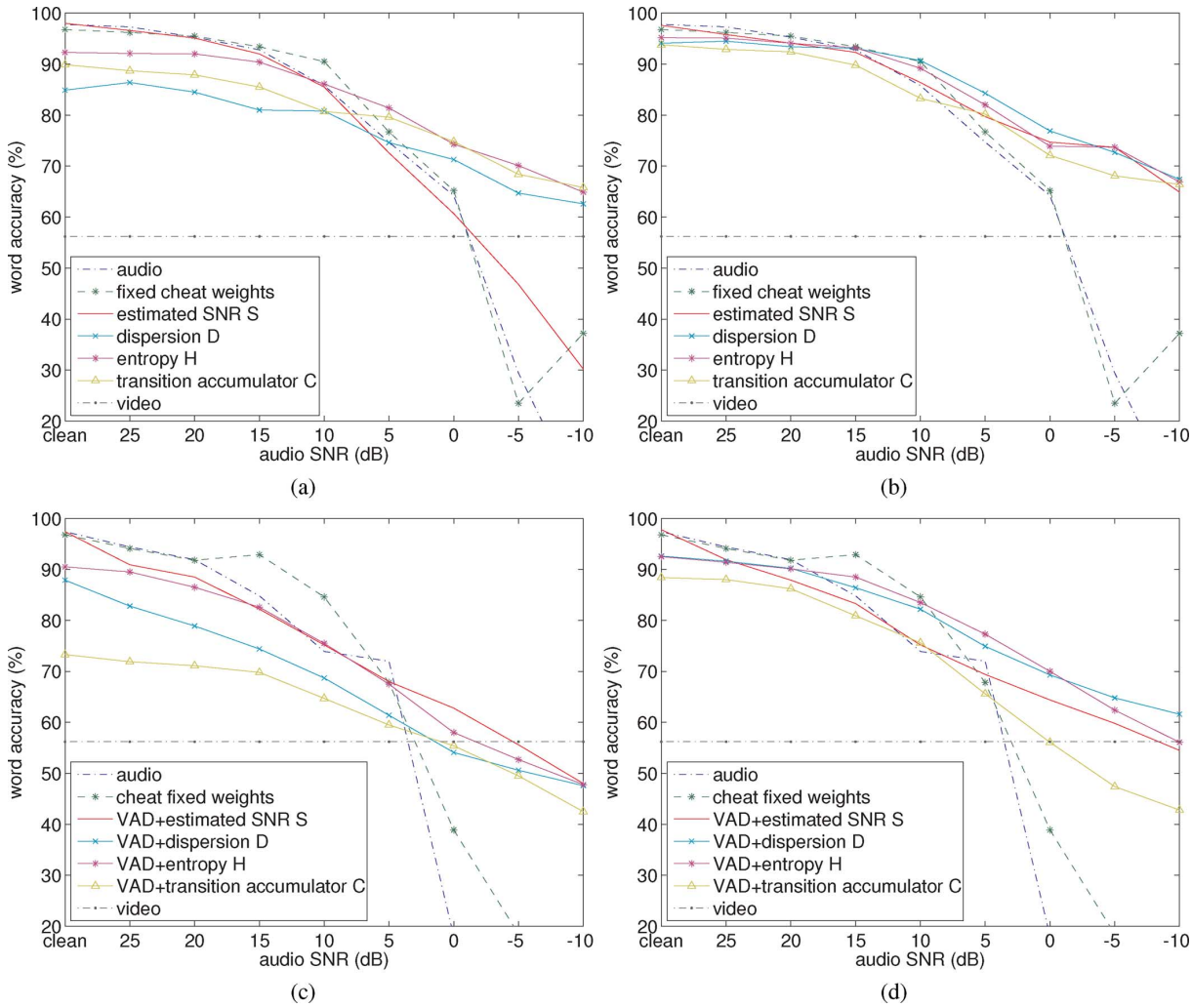

Fig. 7. Performance of cheat fixed weights HMM-ANN system and dynamic weights obtained with one and two mappings. Comparison of (a) with (b) and (c) with (d) show the improvement obtained by the inclusion of the VAD in the ANN system. In (b) and (d), the poor performance of the "cheat" fixed weights in low SNR conditions shows the necessity of including the detection of silences in the weighting strategy. (a) One mapping, audio with white noise. (b) Two mappings, audio with white noise. (c) One mapping, audio with babble noise. (d) Two mappings, audio with babble noise.

used for the classification of speech and silences. The improvement is more remarkable in the SNR estimator (with p_e over 0.9 for the different SNR levels), whose estimation is based on the correct detection of speech and silence intervals, while the entropy, dispersion and transition accumulator benefit less from the inclusion of a VAD in the weighting system (p_e between 0.7 and 0.8) as they already convey information about the confidence that should be given to the classifier during silence intervals. The gain is also clearer for low SNR levels, when the use of only one mapping mainly shifts the weights to use the modality better suited to the silence detection. In this case, the use of two mappings relies on the VAD to detect silence and speech intervals and then uses the corresponding speech or silence mapping for each confidence measure. As a result, for low SNR the dynamic weights even outperform the fixed cheating weights, which do not consider the fact that the proper detection of silences might require a different weight than the speech intervals.

Using only one mapping in the HMM-ANN system, different measures seem to obtain better results for different working conditions. The SNR estimator performs well for high and medium SNR values, while the measures based on the classifiers confidence gain in very noisy environments. In the GMM system, in its turn, the entropy and transition accumulator do slightly better than the other confidence measures. When the VAD is included in the weighting strategy, however, the different confidence measures perform similarly and the differences in performance are not statistically significant (in terms of p_e) for any of the measures.

To summarize, we see that the improvement obtained by the inclusion of the VAD in the weighting strategy is more relevant than any differences in performance between the confidence measures. Without the use of a VAD, different confidence measures obtain better performances for different systems and levels of noise, while the introduction of a VAD into the weighting system improves the performance of all the confidence measures and leads to statistically equivalent results for the different measures. In this case, the proposed transition accumulator ${f_\mathcal{C}}^{\mathrm{VAD}}$ performs equivalently to other classifier's derived measures ${f_\mathcal{H}}^{\mathrm{VAD}}$ or ${f_\mathcal{D}}^{\mathrm{VAD}}$ and is computationally simpler. Similarly, the estimated SNR measure ${f_\mathcal{S}}^{\mathrm{VAD}}$ provides good re-

sults and is easier to compute than the entropy or dispersion of emission likelihoods.

VI. Summary and Conclusion

We presented our work on stream weighting for AV-ASR systems, where weights are introduced to control the contribution of each stream to the recognition task. We focus on the use of dynamic weights in changing environmental conditions, defining the value of the weights as function of different measures of the confidence associated to the stream. The main contributions of the paper are the following: the experimental investigation of dynamic weighting schemes in different noisy environments and system architectures, experimental proof of the effectiveness of introducing a VAD in the weighting scheme and the proposal of a new confidence measure computationally simpler than entropy or dispersion of log-likelihoods.

Based on the signal itself we estimate the SNR present on the audio channel, while we measure the classifier's confidence associated to the stream in terms of the dispersion and the entropy of the class probability distributions. We show how each measure is implemented and suits HMM-ANNs or HMM-GMMs systems and propose a new measure based on the transition probabilities common to both HMM architectures. Evaluating the different stream confidence measures and taking into account the classifiers behavior for the different speech classes, we improve recognition results by the introduction of different mappings for the speech and silence classes.

Experimental results show that dynamic weights perform well in a variety of conditions. For high and medium SNRs, a weighting algorithm based on the classifier's reliability estimators performs well. For very noisy environments, however, the confusion with the silence class is the main cause of failure of the systems and the weighting should first avoid the confusions with the silence class and then focus on recognition of speech. In fact, statistical analysis of the results show that the increase in performance associated to differentiating between silence and speech on the definition of the stream weights is more relevant than any difference in performance between the different reliability measures.

References

[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds. Cambridge, MA: MIT Press, 2004.

[2] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.

[3] N. Fox, R. Gross, J. Cohn, and R. Reilly, "Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts," *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 701–714, Jun. 2007.

[4] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Commun.*, vol. 34, no. 1-2, pp. 25–40, 2001.

[5] G. Potamianos and H. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1998, pp. 3733–3736.

[6] S. Nakamura, H. Ito, and K. Shikano, "Stream weight optimization of speech and lip image sequence for audio-visual speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, 2000, vol. III, pp. 20–23.

[7] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 853–856.

[8] C. Miyajima, K. Tokuda, and T. Kitamura, "Audio-visual speech recognition using MCE-based HMMs and model-dependent stream weights," in *Proc. Int. Conf. Spoken Lang. Process.*, 2000, vol. II, pp. 1023–1026.

[9] E. Sánchez-Soto, A. Potamianos, and K. Daoudi, "Unsupervised stream-weights computation in classification and recognition tasks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 3, pp. 436–445, Mar. 2009.

[10] A. Potamianos, E. Sanchez-Soto, and K. Daoudi, "Stream weight computation for multi-stream classifiers," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. 1, pp. 353–356.

[11] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 1, pp. 857–860.

[12] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 468–472.

[13] A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Humans and Machines*, D. Stork and M. Hennecke, Eds. New York: Springer, 1996, pp. 461–471.

[14] S. Cox, I. Matthews, and A. Bangham, "Combining noise compensation with visual information in speech recognition," in *Proc. Eur. Tutorial Workshop Audio-Visual Speech Process.*, 1997.

[15] S. Gurbuz, Z. Tufekci, E. Patterson, and J. Gowdy, "Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 177–180.

[16] P. Teissier, J. Robert-Ribes, and J. Schwartz, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 629–642, Nov. 1999.

[17] U. Meier, W. Hurst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1996, pp. 833–836.

[18] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 173–176.

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

[20] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, 2000.

[21] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

[22] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, vol. 2002, pp. 1260–1273, 2002.

[23] E. Marcheret, V. Libal, and G. Potamianos, "Dynamic stream weight modeling for audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 4.

[24] A. Garg, G. Potamianos, C. Neti, and T. Huang, "Frame-dependent multi-stream reliability indicators for audio-visual speech recognition," in *Proc. Int. Conf. Multimedia Expo.*, 2003, pp. 605–608.

[25] L. Rabiner and B. Juang, "An introduction to Hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[26] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993, Signal Processing Series.

[27] C. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford Univ. Press, 1995.

[28] N. Morgan and H. Bourlard, "Continuous speech recognition, an introduction to the hybrid HMM/connectionist approach," *IEEE Signal Process. Mag.*, vol. 12, no. 3, pp. 25–42, May 1995.

[29] A. Ganapathiraju, J. Hamaker, and J. Picone, "Hybrid SVM/HMM architectures for speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, 2000, vol. 4, pp. 504–507.

[30] J. Movellan and G. Chadderdon, "Channel separability in the audio-visual integration of speech: A Bayesian approach," *Nato ASI Series F Comput. Syst. Sci.*, vol. 150, pp. 473–488, 1996.

[31] D. Massaro and D. Stork, "Speech recognition and sensory integration," *Amer. Sci.*, vol. 86, no. 3, pp. 236–244, 1998.
[32] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.
[33] K. Kirchhoff and J. Bilmes, "Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1999, pp. 693–696.
[34] F. Berthommier and H. Glotin, "A new SNR-feature mapping for robust multistream speech recognition," in *Proc. Int. Congr. Phon. Sci.*, 1999, pp. 711–715.
[35] L. Terry, D. Shiell, and A. Katsaggelos, "Feature space video stream consistency estimation for dynamic stream weighting in audio-visual speech recognition," in *Proc. Int. Conf. Image Process.*, 2008, pp. 1316–1319.
[36] X. Shao and J. Barker, "Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment," *Speech Commun.*, vol. 50, no. 4, pp. 337–353, 2008.
[37] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[38] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2003, pp. 741–744.
[39] J. Moré and D. Sorensen, "Computing a trust region step," *SIAM J. Sci. Statist. Comput.*, vol. 4, p. 553, 1983.
[40] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Moving-talker, speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 11, pp. 1189–1201, 2002.
[41] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Tech. Rep DRA Speech Research Unit, Malvern, U.K., 1992.
[42] E. Lombard, "Le signe de l'élevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 101–119, p. 25, 1911.
[43] M. Gurban and J.-P. Thiran, "Information theoretic feature extraction for audio-visual speech recognition," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4765–4776, Dec. 2009.
[44] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, 2003.
[45] K. Livescu *et al.*, "Articulatory feature-based methods for acoustic and audio-visual speech recognition," in *Proc. Final Workshop Report, Center for Lang. Speech Process., John Hopkins Univ.*, 2006, vol. 4.
[46] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, U.K.: Entropic Ltd., 1999.
[47] D. Dean, P. Lucey, S. Sridharan, and T. Wark, "Weighting and normalization of synchronous HMMs for audio-visual speech recognition," in *Int. Conf. Auditory-Visual Speech Process.*, 2007, pp. 110–115.
[48] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *Proc. Eur. Signal Process. Conf.*, 2008.
[49] S. Takeuchi, T. Hashiba, S. Tamura, and S. Hayamizu, "Voice activity detection based on fusion of audio and visual information," in *Proc. Int. Conf. Audio-Visual Speech Process.*, 2009, pp. 151–154.
[50] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 409–412.

Virginia Estellers (M'10) was born in Barcelona, Spain, in 1984. She received the M.Sc. degree in mathematics and electrical engineering from Universitat Politècnica de Catalunya, Barcelona, Spain, in 2008. She is currently pursuing the Ph.D. degree at the Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

Since September 2008, she has been a Research and Teaching Assistant at the Signal Processing Laboratory, EPFL. Her current research interests include PDE and variational models for image processing.

Mihai Gurban was born in Timisoara, Romania, in 1979. He received the Computer Science engineer diploma from the Politehnica University, Timisoara, Romania, in 2003 and the Ph.D. degree from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 2009.

After spending a year as a Postdoctoral Research Fellow at EPFL, at the Signal Processing Laboratory (LTS) , he is now working in the industry. His scientific interests include multimodal signal processing, dimensionality reduction, image processing, speech recognition, and machine learning

Jean-Philippe Thiran (M'93-SM'05) was born in Namur, Belgium, in 1970. He received the Elect. Eng. and Ph.D. degrees from the Universite Catholique de Louvain (UCL), Louvain-la-Neuve, Belgium, in 1993 and 1997, respectively.

He joined the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in February 1998 as a Senior Lecturer, responsible for the Image Analysis Group. Since 2004, he has been the Director of the EPFL Signal Processing Lab (LTS5) and in 2011 he was promoted to Associate Professor of signal processing at EPFL. He also holds a part-time associate professor position at the University of Lausanne School of Medicine. His current scientific interests include image segmentation, prior knowledge integration in image analysis, partial differential equations and variational methods in image analysis, multimodal signal processing, medical image analysis, including multimodal image registration, segmentation, computer-assisted surgery, and diffusion MRI. He is author or coauthor of one book, nine book chapters, 90 journal papers, and more than 150 peer-reviewed papers published in proceedings of international conferences. He holds four international patents.

Dr. Thiran was Co-Editor-in-Chief of the *Signal Processing* journal (published by Elsevier Science) from 2001 to 2005. He is currently an Associate Editor of the *International Journal of Image and Video Processing* and member of the Editorial Board of *Signal, Image, and Video Processing* (both published by Springer). He is currently an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. Among many other scientific duties, he was the General Chairman of the 2008 European Signal Processing Conference (EUSIPCO 2008), the tutorial co-chair of the IEEE International Conference on Image Processing (ICIP) in 2011 and will be the technical co-chair of ICIP 2015.