WEIGHTING SCHEMES FOR AUDIO-VISUAL FUSION IN SPEECH RECOGNITION

Hervé Glotin[1], *Dimitra Vergyri*[2], *Chalapathy Neti*[3], *Gerasimos Potamianos*[3], *Juergen Luettin*[4]

[1] ICP Grenoble, France & IDIAP Martigny, Switzerland
[2] SRI International, Menlo Park, CA 94025, USA
[3] IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
[4] Ascom Systec AG, 5506 Maegenwil, Switzerland

ABSTRACT

In this work we demonstrate an improvement in the state-of-the-art large vocabulary continuous speech recognition (LVCSR) performance, under clean and noisy conditions, by the use of visual information, in addition to the traditional audio one. We take a *decision fusion* approach for the audio-visual information, where the single-modality (audio- and visual- only) HMM classifiers are combined to recognize audio-visual speech. More specifically, we tackle the problem of estimating the appropriate combination weights for each of the modalities. Two different techniques are described: The first uses an automatically extracted estimate of the audio stream reliability in order to modify the weights for each modality (both clean and noisy audio results are reported), while the second is a discriminative model combination approach where weights on pre-defined model classes are optimized to minimize WER (clean audio only results).

1. INTRODUCTION

In [1] we present decision fusion algorithms that focus on both state synchronous (combining likelihoods at the state level) and phone synchronous modeling (combining likelihoods at the phone level) of the audio and visual streams. The model investigated in this approach was a multi-stream HMM, and a phone synchronous variant, a product HMM. The state (phone) conditional observation likelihood of these models is the product of the observation likelihoods of their audio-only and visual-only stream components, raised to appropriate stream exponents that capture the reliability of each modality.

In this article, we expand on exponent estimation further. Two techniques are presented:

In the first technique we investigate possible refinements of stream exponent dependence, making use of an automatically extracted audio stream reliability, estimated on the basis of the degree of voicing present in the audio signal [2]. This approach follows the concept of audio-visual adaptive weights used in [3]. We first consider exponents that are utterance dependent, estimating the average voicing over the utterance frames. This approach gives us a relative improvement of about 7% in clean and 26% in noisy speech (cocktail party "babble" speech noise at 10dB SNR), when compared to the audio only WER. We also show an improvement compared to the multi-stream HMM used in [1]. Next we explore frame dependent exponents, using per frame voicing estimates. The results are compared with the baseline and the per utterance exponent results.

The second technique is a discriminative model combination (DMC) [4] approach: The audio and visual streams are used independently to train models which are then combined along with a language model, with weights optimized to minimize the WER on a held out set. The two modalities are only synchronized at the utterance level in this approach. Similar to the work in [5, 6], we consider exponents static for each modality, or dynamic (hypothesis dependent). We present results only in clean speech conditions, where we show a 5% relative improvement over a baseline which combines two different audio models by N-best rescoring.

Section 2 describes the experimental setup for this work. Section 3 presents the general form of the fusion model. The two weighting techniques are presented in sections 4 and 5 respectively. We conclude with a discussion of the accomplishments of this work and some ideas for future directions.

2. DATABASE - EXPERIMENTAL SETUP

To allow experiments on continuous large vocabulary, speaker independent audio-visual speech recognition, a database has been collected at IBM for the purposes of the CLSP/JHU summer 2000 workshop [7]. For the noisy experiments a cocktail party "babble" speech noise was added to the audio signal at 10dB SNR. Baseline ASR systems were obtained during the workshop, for clean and noisy audio, using HTK. The acoustic models were cross-word triphone HMMs with about 75K Gaussian mixtures. All the models developed at the workshop were used to rescore word lattices, generated with the IBM LVCSR recognizer which used pentaphone cross-word HMMs with about 50K Gaussian mixtures [8]. The speaker independent (SI) test set (1038 utterances; 2.5 hours) defined in the workshop was used for these experiments, while a held out set was used for tuning the parameters.

3. MULTI-STREAM HMMS FOR AUDIO-VISUAL FUSION

In this work we are using a multi-stream HMM to combine the audio and visual modalities (streams). As described in [1] the model computes the class conditional observation likelihood as a product of the observation likelihoods of its single-stream components, raised to the appropriate *stream exponents* that capture the reliability of each modality. Given the bimodal (audio-visual) observation vector $\mathbf{o}^{(t)} = \{\mathbf{o}_{\mathrm{A}}^{(t)}, \mathbf{o}_{\mathrm{V}}^{(t)}\}$ the state emission (class conditional) probability of the multi-stream HMM is:

$$Pr[\mathbf{o}^{(t)}|\mathsf{c}] = \prod_{s \in \{\mathrm{A},\mathrm{V}\}} \Big[\sum_{j=1}^{J_{sc}} w_{scj} \mathcal{N}_{D_s}(\mathbf{o}_s^{(t)}; \mathbf{m}_{scj}, \mathbf{s}_{scj})\Big]^{\lambda_{sct}} \quad (1)$$

where the stream exponents λ_{sct} are non-negative and, in general, depend on the modality s, the HMM state (class) c, and locally, on the utterance frame (time) t. Such model has been considered

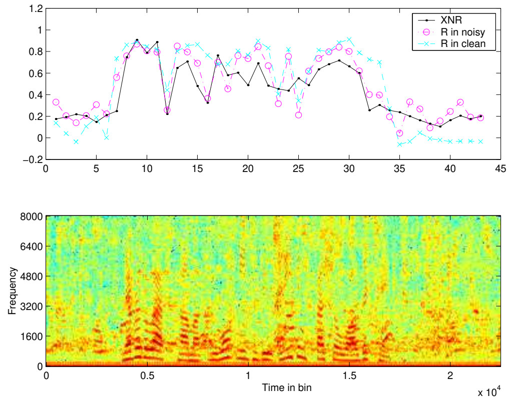

Fig. 1. *Top*: Local estimates of R in clean and in noisy speech ("babble" noise 10dB), and XNR reference, for a database utterance. All calculations are performed on 128ms speech windows shifted by 64ms. *Bottom*: Noisy audio spectrogram of the same utterance.

in multi-band audio-only ASR, among others [9], and, as a two-stream HMM, in small-vocabulary audio-visual ASR tasks [10, 11, 3].

In (1) the likelihoods are combined at the state level (state synchronous multi-stream model). We can choose to combine the likelihoods at a class level (class can be phone, syllable, word, or even utterance) allowing different degrees of asynchrony between the two streams. We refer to this model as a class-synchronous product model [1, 8]. During maximum likelihood (ML) training of the model, the weights λ_{sct} were kept fixed. Their values were optimized on held out data prior to the ML training of the Gaussian mixtures of the single-stream HMMs.

4. VOICING AS A MEASURE OF AUDIO RELIABILITY

Various signal-to-noise ratio estimates have been used in the literature in order to assign an audio-stream weight [12]. Here, we propose the use of a measure of voicing, correlated with SNR, as a means of estimating the reliability of the audio observations, and we apply it to audio-visual weighting. We employ an equivalent *harmonicity index* (HNR) [13, 14] to estimate the average voicing per utterance. Based on this index, we subsequently estimate utterance based stream exponents.

We use the autocorrelogram of a demodulated signal as a basis for differentiating between a harmonic signal and noise. In the case of Gaussian noise, the correlogram of a noisy cell is less modulated than a clean one [13]. The peaks in the autocorrelogram of the demodulated cell isolate the various harmonics in a signal. This can be used to separate a mixture of harmonic noises and a dominant harmonic signal. It is interesting that such separation can be efficiently accomplished, using a time window of duration in the same range as the average phoneme duration.

We get an audio reliability estimate for each frame of 128ms duration. Before the autocorrelation we compute the demodulated signal after half wave rectification, followed by band-pass filtering in the pitch domain ([90,350] Hz). For each cell, we calculate the ratio $R = R1/R0$, where $R1$ is the local maximum in time delay segment corresponding to the fundamental frequency, and $R0$ is the cell energy. This measure is strongly correlated with SNR in the 5–20dB range [15]. Figure 1 demonstrates explicitly the correlation of this measure in clean and noisy speech with the noisy signal SNR: We plot R estimates on 128ms speech windows of a noisy utterance, against the R estimates in the clean audio case, and a linear SNR-alike measure, defined as $\mathrm{XNR} = (S/(S+N))$, where S is the energy of the clean signal and N is the noise only energy. Notice that R and XNR do not give exactly the same kind of information, but they are quite strongly correlated. Their correlation factor is 0.84, computed over the entire SI test set. Locally, R is higher than XNR on voiced parts, and it is lower on other parts. This local divergence on a per frame level could be exploited to obtain a local stream weighting scheme.

	Clean audio WER% (relative)		Noisy audio WER% (relative)	
Audio-only	14.44	—	48.10	—
AV-MS	14.62	+1.2	36.61	-23.9
AV-MS-UTTER	13.47	-6.7	35.27	-26.7
AV-PROD	14.19	-1.7	35.21	-26.8
AV-PROD-UTTER	—		35.43	-26.3
AV-PROD-LOCAL	—		37.15	-22.8

Table 1. Audio-visual decision fusion WER(%). We compare with three different baselines: Audio-only, AV-MS and AV-PROD. Relative WER improvements are computed as the relative (%) gain over the audio-only baseline.

4.1. Utterance Dependent Stream Exponents

In this first approach, audio speech reliability is calculated only from the regions where the speech is dominant. We assume that regions where local SNR is higher than 0dB (strongly correlated to the regions where $\mathrm{R} > 0.5$; see also Figure 1) are speech regions. We subsequently calculate stream exponents $\lambda_{\mathrm{A}\,t}$, constant for all t within the utterance, to be the mean of all R values higher than 0.5. We assume this to be an adequate estimate of voicing within the utterance. Then, in (1): $\lambda_{\mathrm{V}\,t} = 1 - \lambda_{\mathrm{A}\,t}$.

We found that the audio weight $\lambda_{\mathrm{A}\,t}$ is mostly speaker dependent and, in a smaller extent, utterance dependent [8]. For the entire SI test data set, the average $\lambda_{\mathrm{A}\,t}$ is calculated to be 0.79 and 0.73 for the clean and noisy audio case, respectively.

4.2. Local Frame Dependent Stream Exponents

Since we estimate frame reliability on 128ms signal intervals, we can use these estimates to compute the local frame weights, using the piecewise-linear mapping function proposed by Meier in [16]. We experimented also with other linear functions of R inspired from studies as in [11] without success so far. We optimized the parameters of the Meier function on a small held out set. Thus we compute the weights as: $\lambda_{\mathrm{A}\,t} = min(max(R(t), 0.5), 0.7)$.

4.3. Experimental Results - Discussion

The WER results for different systems, for the SI test set, are presented in Table 1. We compare our results to an audio-only and two audio-visual baselines. The audio-only baseline uses the HTK trained acoustic model (see section 2). The AV-MS model is the state synchronous audio-visual multi-stream model described in (1), while the AV-PROD model is a phone synchronous product model (see section 3). The details for the development of these models are described in [1, 8]. These baseline systems are trained using audio weight λ_A optimized on a held out set for clean (noisy) speech: 0.7 (0.6) for AV-MS, and 0.6 (0.7) for AV-PROD. The video weight was set to $\lambda_V = 1 - \lambda_A$. These values for the exponents were kept fixed during testing with these systems.

In both clean and noisy conditions the AV-MS-UTTER model (utterance dependent stream exponents) outperformed the comparable AV-MS one, resulting to almost 7% relative WER reduction with respect to the audio-only system. In the clean audio case this model even outperformed the AV-PROD model.

Our attempts to modify the weights for the AV-PROD model in noisy conditions were not as successful. The AV-PROD-UTTER model is slightly worse that the AV-PROD (fixed global weights). The result deteriorates more when we use local per frame weights (AV-PROD-LOCAL). It seems that the ML training of the phone-synchronous product model, captures some of the information about stream reliability that we are trying to use in order to modify the original weights. We notice that our improvement over the model with global weights is much better in clean than in noisy conditions. This may be due to the fact that our estimate of the audio stream reliability (voicing) is more accurate in clean speech.

Further investigation is due to examine more appropriate variable weights over the noisy product model. A different mapping function for R can be explored for that purpose [15].

5. DISCRIMINATIVE COMBINATION OF AUDIO AND VISUAL MODELS

The Discriminative Model Combination (DMC) approach [4] aims at an optimal integration of independent sources of information in a log-linear model that computes the probability for a hypothesis. The parameters of this new model are the weights of the log-linear combination, and are optimized to minimize the errors in a held out set.

When we have independent observation streams as sources of information, and we have trained maximum likelihood models independently for each of these streams, then the DMC approach is equivalent to optimizing the stream weights for model (1). In the implementation of this approach though, we use an asynchronous version of that model allowing the two streams to be synchronized only at the utterance boundaries.

The combination of the models can be performed either *statically*, with constant weights [4], or *dynamically*, where the parameters may vary for different segments of a hypothesis [17, 6]. In the dynamic combination the weights aim to capture the dynamic change of confidence on each of the models combined for each hypothesized segment.

5.1. Static Combination

We can combine the audio and visual models' scores, along with a language model score, as independent sources of information in the DMC framework. If we denote by $P_s(\mathbf{h}|\mathbf{O}_s)$, with $s \in \{\mathrm{A}, \mathrm{V}\}$, the probability provided by the audio (visual) models [1], we define, in the DMC framework, the log-linear model that combines all the available information $\mathcal{I}$ (audio/visual/linguistic information) as:

$$P(\mathbf{h}|\mathcal{I}) = \frac{1}{Z_\Lambda(\mathcal{I})} \left(\prod_{s \in \{\mathrm{A},\mathrm{V}\}} P_s(\mathbf{h}|\mathbf{O}_s)^{\lambda_s} \right) P_{LM}(\mathbf{h})^{\lambda_{LM}} \qquad (2)$$

where $P_{LM}(\mathbf{h})$ is the language model probability and $Z_\Lambda(\mathcal{I})$ is a normalization factor so that the probabilities for all $\mathbf{h} \in \mathcal{H}$ add to one. In this formulation we only have one static weight for each stream.

5.2. Dynamic Combination - Phone Dependent Weights

We can combine the scores from the available information sources dynamically, within the simple form of an exponential model, by weighting each of the scores with different exponents, for different segments of a hypothesis [17]:

$$P(\mathbf{h}|\mathcal{I}) = \frac{1}{Z_\Lambda(\mathcal{I})} \left(\prod_{i=1}^{N} \prod_{s \in \{\mathrm{A},\mathrm{V}\}} P_s(h_i)^{\lambda_s(h_i)} \right) P_{LM}(\mathbf{h})^{\lambda_{LM}} \qquad (3)$$

where h_i is the ith segment in hypothesis $\mathbf{h}$. In this model we see that the exponent value changes with time across each utterance.

The weights $\lambda(\cdot)$ we use are tied across different classes of segments so that we have only a small number of parameters to optimize. Motivated by the work for multi-lingual model combination [5, 6], we chose the stream weights to depend on the identity of the hypothesized phones. Since the phones are not well defined for the visual model, we used visemic classes instead (these are the visually distinctive phones [8]).

5.3. Parameter Optimization

The above defined model is used to rescore the N-best lists and choose the MAP candidate. We train the parameters $\lambda(\cdot)$ in (2) and (3) so that the empirical word error count induced by the model is minimized. Since the objective function is not smooth, gradient descend techniques are not appropriate for estimation. We use the simplex downhill method, known as amoeba search to minimize the number of word errors on a held out set [17].

5.4. Experimental Results - Discussion

We used only the clean speech utterances for our experiments. A held out set of about 1500 utterances was set aside in order to optimize the weights, and the SI test set available at the workshop was used for testing. For the purposes of the experiments, 2000-best hypotheses were obtained for each utterance using acoustic model scores provided by IBM and they were then rescored with the new acoustic and visual models created during the workshop using HTK [2].

[1] We note that the acoustic model and visual models typically provide a conditional probability of the observations given the hypothesis but we approximate the likelihood of the hypothesis, given the observations, using Bayes rule and a uniform language model $P_u(\mathbf{h}) = \mathbf{c}$. Thus:

$$P_s(\mathbf{h}|\mathbf{O}_s) = \frac{\tilde{P}(\mathbf{O}_s|\mathbf{h})P_u(\mathbf{h})}{\sum_{\mathbf{h}'} \tilde{P}(\mathbf{O}_s|\mathbf{h}')P_u(\mathbf{h}')} = \frac{\tilde{P}(\mathbf{O}_s|\mathbf{h})}{\sum_{\mathbf{h}'} \tilde{P}(\mathbf{O}_s|\mathbf{h}')} = \tilde{P}(\mathbf{O}_s|\mathbf{h})$$

[2] The IBM system used to generate the N-best lists had a WER of 14.24%. Due to the rescoring of these N-best hypotheses with the HTK audio only model, the new baseline is better than the one obtained using the HTK model alone in Table 1 (ROVER effect).

Experiment	train WER	SI test WER (relative)
0. Baseline acoustic [2]	12.8	13.65 (–)
1. Static (acoustic + visual) weights	12.5	13.35 (-2.0)
2. 1 acoustic + 13 visemic weights	12.2	13.22 (-3.1)
3. phonemic + 13 visemic weights	11.8	12.95 (-5.1)

Table 2. DMC experimental results on clean audio.

We performed 3 experiments:
Experiment 1: The audio and visual models are combined statically with one weight for each of the models.
Experiment 2: One global weight is still used for the audio model scores, but we use 13 different weights for visual models corresponding to the each of the 13 visemic classes.
Experiment 3: Different weights are used for each of the 43 audio phone-models and each of the 13 visemic-classes.

The results are depicted in Table 2. We found this method efficient enough to obtain an extra 5% relative improvement over the improved baseline, resulting to a total of 10% relative improvement over the workshop clean audio baseline (see Table 1). Therefore half of the improvement in the system is due to the use of the visual information.

Our exponent parameterization scheme is rather simple: We only allow exponents that depend on the identity of hypothesized phones/visemes. Different exponent classification schemes, using information about the reliability of each model, might be worth exploring in future work. We also need to point out one of the limitations of this approach: The lack of synchronization. The two streams are used independently and their scores are combined only at the utterance level. This way the weaker visual model cannot utilize the information provided by its better audio model about the word or phone boundaries.

6. CONCLUSIONS

We demonstrated two different techniques for improving speech recognition using audio-visual models. The techniques are aiming at optimal weighting schemes for audio-visual fusion. We showed significant improvements in LVCSR, particularly in clean speech which in previous work had been difficult to improve with the use of visual information.

Future work can further explore HNR based exponent modification that could eventually show improvements even in the product model. It will also be interesting to examine ways of combining the two approaches. Namely, we can explore the discriminative optimization of HNR-based class exponents, and the application of DMC with different levels of asynchrony between the two streams.

7. ACKNOWLEDGMENTS

H. Glotin is funded by EEC projects TMR SPHERE and LTR RESPITE. The authors also wish to thank the organizers of the CLSP/JHU summer'00 workshop and all the participants of the AVSR project group.

8. REFERENCES

[1] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *Proc. ICASSP*, 2001, submitted.

[2] F. Berthommier and H. Glotin, "A measure of speech and pitch reliability from voicing," in *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, F. Klassner, Ed., 1999, Computational Auditory Scene Analysis (CASA) workshop, pp. 61–70.

[3] A. Rogozan, P. Deléglise, and M. Alissali, "Adaptive determination of audio and visual weights for automatic speech recognition," in *Proc. European Tutorial Workshop on Audio-Visual Speech Processing (AVSP)*, 1997, pp. 61–64.

[4] P. Beyerlein, "Discriminative model combination," in *Proc. ICASSP*, 1998, vol. 1, pp. 481–484.

[5] P. Beyerlein, W. Byrne, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Pterek, J. Picone, D. Vergyri, and W. Wang, "Towards language independent acoustic modeling," in *Proc. ICASSP*, 2000.

[6] D. Vergyri., S. Tsakalidis., and W. Byrne, "Minimum risk acoustic clustering for multilingual acoustic model combination," in *Proc. ICSLP*, 2000, pp. 873–876.

[7] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Proc. ICASSP*, 2001, submitted.

[8] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Tech. Rep., CLSP/Johns Hopkins University, Baltimore, 2000, available online at http://www.clsp.jhu.edu/ws2000/final_reports/avsr/.

[9] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. ICSLP*, 1996, vol. 1, pp. 426–429.

[10] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. ICASSP*, 1998, vol. 6, pp. 3733–3736.

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

[12] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines: Models, Systems and Applications*, D. G. Stork and M. E. Hennecke, Eds., pp. 331–349. Springer, Berlin, 1996.

[13] F. Berthommier and H. Glotin, "A new SNR-feature mapping for robust multistream speech recognition," in *Proc. International Congress on Phonetic Sciences (ICPhS)*, 1999, vol. 1, pp. 711–715.

[14] E. Yumoto, W. J. Gould, and T. Baer, "Harmonic to noise ratio as an index of the degree of hoarseness," *Journal of the Acoustical Society of America*, vol. 1971, pp. 1544–1550, 1982.

[15] H. Glotin and F. Berthommier, "Test of several external posterior weighting functions for multiband full combination asr," in *Proc. ICSLP*, 2000, vol. I, pp. 333–336.

[16] U. Meier, W. Hurst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading," in *Proc. ICASSP*, 1996.

[17] D. Vergyri, *Integration of Multiple Knowledge Sources in Speech Recognition Using Minimum Error Training*, Ph.D. Thesis, CLSP/Johns Hopkins University, Baltimore, 2000.