Sensory Integration in Audiovisual Automatic Speech Recognition

Peter L. Silsbee
Dept. of Electrical and Computer Engineering
Old Dominion University
Norfolk, VA 23529

Abstract

Methods of integrating audio and visual information in an audiovisual HMM-based ASR system are investigated. Experiments involve discrimination of a set of 22 consonants, with various integration strategies. The role of the visual subsystem is varied; for example, in one run, the subsystem attempts to classify all 22 consonants, while in other runs it attempts only broader classifications.

In a second experiment, a new HMM formulation is employed, which incorporates the integration into the HMM at a pre-categorical stage. A single variable parameter allows the relative contribution of audio and visual information to be controlled. This form of integration can be very easily incorporated into existing audio-based continuous speech recognizers.

SNR	Number of Misclassifications					
	Place		Manner		Voicing	
	audio	AV	audio	AV	audio	AV
25	2	1	3	2	1	2
15	7	3	7	5	1	0
5	20	5	18	7	2	6

Table 1: Results of an experiment from [1]. Consonant misclassifications were categorized according to whether the correct class differed from the chosen class in place of articulation, manner, and/or voicing, for audio and audiovisual (AV) ASR. Even though audiovisual ASR performed better than audio ASR for overall classification, classification with respect to voicing was adversely affected in some cases.

1 Introduction

Audiovisual methods of automatic speech recognition (ASR) have recently gained attention as they offer improved robustness and accuracy for ASR. Recently reported work includes [1, 2, 3]. In order to maximize the performance of these systems, several issues must be addressed. First, clearly, the new visual methods should be designed so that they are as accurate as possible when applied by themselves. Second, methods should be developed which extend the successful techniques used for audio ASR, such as hidden Markov models, to visual ASR. Third, an effective *integration* strategy must be developed for fusing the two information. This paper addresses the third issue.

It is evident from previous studies, both in ASR and in human speech perception, that the two sensory modalities have different strengths and weaknesses, and in fact that to a large extent these complement each other. The most striking example of this comes from classification of consonants. When consonants are categorized according to their *place* of articulation, *manner* of articulation, and *voicing*, audio-based recognition performs the least well at classification with respect to place, and the best respect to voicing. This is the exact opposite of what is found for visual recognition. This effect has been well documented for human speech perception [4], and seems to hold for many ASR systems as well [3, 1]. An example is shown in Table 1. In this example, overall recognition performance of an ASR system improved when lipreading was incorporated; however, *more errors occurred with respect to classification of voicing.* If this is truly an inherent part of audiovisual ASR, then an optimal integration strategy should account for this. On the other hand, if this is merely a human limitation, then ways may be found to overcome it.

In many situations, the quality of both signals is subject to variation. Differences in microphone characteristics, acoustic environment, speaker-to-microphone distance, etc. are all well known factors which degrade the performance of audio ASR systems. Visual recognizers face an even more daunting array of variables, ranging from lighting conditions to camera distance to the fact that the speaker may turn his or her head completely away from the camera. Quality estimates

1058-6393/95 $4.00

of both signals should be used to guide the integration process. This would allow more reliance on the visual signal when acoustic noise conditions are bad, for example. The issue of obtaining these quality estimates is an important one; however, it is not explicitly addressed in this paper. Instead, we focus on methods which allow a simple adjustment of the relative importance of each information stream.

It should be noted that a statistical classifier, optimally trained, will automatically give appropriate weights to the individual pieces of information. However, there are at least two reasons why this should not be relied upon. First, as mentioned above, conditions are variable; thus, it is not always possible to assume that a classifier is optimally trained for the conditions under which the system is used. Second, at least for the present, most researchers in audiovisual ASR have access only to a very limited set of training data. Therefore, most current audiovisual ASR systems are undertrained.

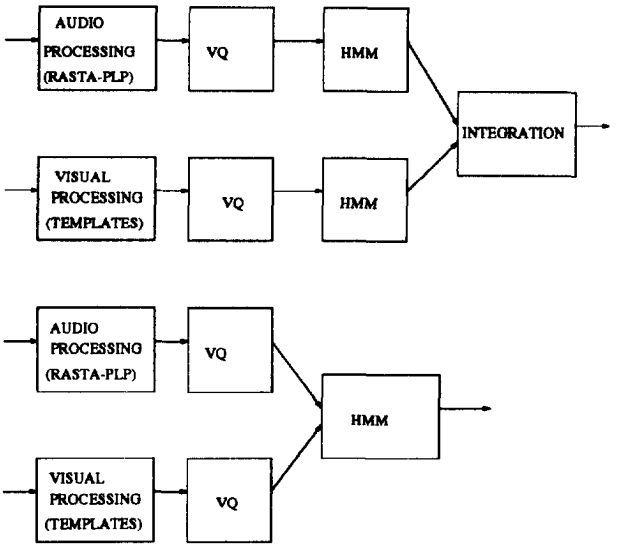

Figure 1: Block diagram of the general audiovisual speech recognition system, showing the two generic integration integration schemes.

2 The Recognizer

The experiments described in this paper were performed on a version of the LEAPS (Lipreading to Enhance Automatic Perception of Speech) system, which has been described more fully in [1]. The original version of LEAPS consisted of independent audio and visual processors with independent HMM-based scoring modules. The audio processor used Perceptual Linear Prediction [5] and vector quantization, while the visual processor used a modified form of VQ which was applied directly to the images. Integration was performed by additive combination of the audio and visual scores:

$$S(c) = \lambda S_a(c) + (1 - \lambda) S_v(c) \ , \qquad (1)$$

where c is a class, $S(c)$ is the overall score, and $S_a(c)$ and $S_v(c)$ are the audio and visual scores, respectively. Note that, since HMM scores are log probabilities, that setting $\lambda = 0.5$ corresponds to estimating the overall probability of a class with an assumption of independence. It does *not*, however, necessarily correspond to an equal contribution of information.

The current system is similar but has been modified in several ways. First, the audio processing now uses a low-order mixed RASTA-PLP speech processing [6] subsystem. Secondly, whereas the original LEAPS visual processor used a speaker-dependent VQ which acted directly on image frames, the current version uses a simple deformable model (described in [7]), whose parameters after convergence are vector quantized. For the experiments reported here, neither subsystem currently performs as well as the original subsystems reported in [1]. In the case of the audio subsystem, the performance has been deliberately impaired (through the use of a very low-order representation) so that more meaningful statistics can be obtained (it is difficult meaningfully to measure improvement when the audio subsystem makes only three or four errors). In the case of the visual subsystem, the degraded performance is due in part to the move to a speaker independent analysis method, but largely to the simplicity of the current model, which is under development. The other modifications to the system involve the integration of the audio and visual information, and are the subject of this paper. A block diagram of the system is shown in Fig. 1. This figure also indicates the stages at which integration occurs in different experiments.

3 Integration Strategies

This paper considers two generic integration strategies; namely, integration as an integral part of HMM scoring and integration after HMM scoring. The latter strategy is easy to implement, especially for short utterances (e.g. eq. (1)). It may also be implemented via an N-best rescoring technique [8].

However, integration as an integral part of HMM scoring offers certain advantages. In particular, if Viterbi scoring is used, the most likely state sequence

will be jointly determined by the audio and visual information, and symbol emission probabilities for both a function of that same state sequence in both cases. This is in contrast to a post-HMM integration where the two independent HMM recognizers could base their scores for a particular phoneme on completely different state sequences. From a practical point of view, integration within the Viterbi scoring process also allows the use of the many tools (search algorithms, grammars, etc.) which have been developed for HMM-based recognition.

Another aspect of integration which is explored is the role of the visual recognizer. Although it is widely accepted that there is a fundamental distinguishable unit of visual speech (the viseme) which is used by human lipreaders, there is some disagreement about exactly which sets of phonemes constitute a viseme. It might seem to follow (assuming that a "correct" set of visemes could be identified) that a machine recognition system should operate using these same fundamental units. However, there is no reason to assume that human lipreading performance represents an upper limit on machine performance. In fact, Finn and Montgomery [9] found that a machine could distinguish, based on visual information only, between phonemes which normally would be considered part of the same viseme class. Their results may well not generalize to more practical audiovisual ASR systems; but the possibility, at least, that a machine recognition system could utilize a larger effective set of visemes, should be acknowledged.

With this in mind, three post-HMM integration strategies, all using Eq. (1), are compared. The first operates on the assumption that the machine can distinguish between all phonemes. The second, based on previous results such as those presented in Table 1, which indicate that lipreading can adversely affect voicing decisions, uses a set of visemes such that the visual recognizer plays no part in that decision. Each voiced/unvoiced pair of consonants is merged into a single viseme; for example, /p/ and /b/ form a single class. This results in 14 separate classes out of the original 22. The third strategy uses a more restrictive viseme set, which is given in Table 2. In these latter cases, the quantity $S_v(c)$ in eq. (1), for a given class c, is replaced by the score for the "superclass" to which c belongs. Thus, for example, in the second strategy, the total score for /p/ is calculated from the audio score for /p/, and the visual score for "bilabial stop."

It is worth noting that with an optimally trained statistical classifier, there would be no reason to assume any restrictions on the capabilities of the visual recognizer.

/p/, /b/, /m/	/f/, /v/
/k/, /g/, /ng/	/dh/, /th/
/t/, /d/, /n/, /l/, /s/, /z/	/r/
/sh/, /tsh/, /zh/, /dzh/	/h/

Table 2: Visemes for visual classification. After [10].

Also investigated is a new method for integration which is directly incorporated into the HMM scorer. The description which follows is for a discrete HMM recognizer; however, it can easily be extended to continuous HMMs.

For multiple-codebook discrete HMMs, at each time step, an estimate is made of the probability that the observed symbols would have been emitted by the model, given that the Markov process is in a particular state S_i. Suppose that the three simultaneous audio symbols A_{1i}, A_{2j}, and A_{3k} have been observed (the double subscripts indicate that each symbol has been drawn from a different alphabet). The usual approach is to assume that each symbol is independent, which gives

$$\Pr(A_{1i}, A_{2j}, A_{3k}) = \Pr(A_{1i}) \Pr(A_{2j}) \Pr(A_{3k}) \quad (2)$$

where the conditioning on state S_i has been omitted for simplicity of notation. Thus, the probability of the joint observation is easily calculated (since this is usually carried out in logarithmic domain, these quantities are simply summed). The audiovisual case can be considered as a straightforward extension of the multiple-codebook HMM, where the visual processor simply contributes one more symbol, derived from one more codebook. In the case of an optimally trained recognizer, indeed, this would result in an appropriate estimate of the probability. However, to allow for variability and changes in the relative importance of the two information sources, it is proposed to include a parameter γ which controls their relative influence. If the observed visual symbol is V_l, $\mathbf{O}$ represents the full vector of observations, and N is the size of the visual VQ symbol alphabet, we can estimate:

$$\Pr(\mathbf{O}) \approx \frac{\Pr(A_{1i}) \Pr(A_{2j}) \Pr(A_{3k}) \Pr^{\gamma}(V_l)}{\sum_{n=1}^{N} \Pr^{\gamma}(V_n)} . \quad (3)$$

The normalizing factor $\sum_{n=1}^{N} \Pr^{\gamma}(V_n)$ is required to make the new quantities behave like probabilities. Although it is not indicated by the notation, in general this quantity is a function of the state only, and can be precomputed after training. Of course, the scores derived using the above expression will no longer strictly be probability estimates.

4 Experiments

The experimental data are described in detail in [1]. Ten audiovisual sequences were obtained from a single speaker for each of twenty-two consonants. The sequences ranged from about one-half second to one second in length. The consonants were each presented in an identical /a-C-a/ context; the sequences include this entire utterance. The visual data consisted of sequences of about 15 to 30 80×80 pixel frames, acquired at 30 frames per second. The audio data was sampled at 16000 samples per second, with 12 bits per sample. Six tokens of each consonant were used for training, and four for testing (a total of 88 test tokens).

Performance is measured by the number of misclassifications in each of three articulatory categories (place, manner, and voicing), as well as overall misclassifications and by the average rank of the correct class in the list of scores produced by the classifier (an average rank of 1 corresponds to perfect performance). The average rank criterion measures the seriousness of missed classifications. In the experiments described here, the objective has not been to maximize the overall performance of the system; but strictly to investigate and compare different integration strategies. Note that the visual recognizer by itself currently recognizes only about 25% of the tokens correctly; this can certainly be improved enormously.

In experiment I, the three post-HMM strategies are compared. Table 3 shows the number of misclassifications and the average rank of the correct answer for the audio subsystem and for the audiovisual system for each integration strategy. A value of $\lambda = 0.35$ was used; this was reported in [1] to be an effective value for a wide range of conditions.

A significant improvement is found, as expected, when audiovisual classification is compared to audio-only classification. A 45% reduction in overall error rate is noted, with most of that reduction attributable to better classification with respect to place of articulation.

Restricting the scope of the visual classifier has a minor effect at best. Although one less error occurred with respect to voicing, the effect is certainly not statistically significant. To the extent that one can draw any conclusions at all, it would appear that most or all of the benefit comes from the first restriction (removing visual influence from the voicing decision), and further restriction to the set of eight visemes is unwarranted. It is quite possible, however, that these effects are dependent on the specific training data which is available.

Experiment II was designed to evaluate the within-HMM integration strategy, for the case where the visual recognizer is expected to distinguish between all 22 classes. Equation (3) was applied during Viterbi estimation of the sequence probabilities. A range of γ values was used. It must be noted that, unfortunately, synchronization information is not available for this set of experimental data. Though the audio and visual data were acquired simultaneously, the "start" and "stop" times were quite independent. The HMM/Viterbi based integration requires synchronization, since the final probabilities depend on a single state path which is presumed to have emitted the audio and the visual symbols. As a partial fix, the visual data have been semi-manually time-warped so that the endpoints of the visual and audio signals, as well as the approximate temporal center of each consonant, are matched. This still probably represents rather poorly synchronized signals; thus, the experimental performance reported for this method should be regarded as pessimistic.

	(a)	(b)	(c)	(d)
Voicing	4	5	5	5
Manner	8	6	6	7
Place	12	8	6	10
Overall	18	12	11	14
Avg. Rank	1.34	1.28	1.23	1.30

Table 3: Summary of results from experiment I (number of misclassifications, and average rank of correct answer). (a) Audio only. (b) Integration under the assumption that the visual system can classify all phonemes. (c) Visual classifier restricted to a 14 viseme set where no voicing distinction is made. (d) Visual classifier further restricted to the 8 viseme set listed in Table 2. Note that the overall number of misclassifications is less than the sum of the categorical misclassifications because several phones had more than one attribute misclassified.

Table 4 shows the results of this experiment. The best performance is very similar to that of experiment I. The effect of varying γ is clearly seen; low values of γ indicate that the audio information was given the most weight, while high values of γ indicate that visual information was given more weight. Performance peaks with γ values near 0.5. Compared to audio-only performance, classification with respect to place of articulation is improved the most of the articulatory features. Somewhat surprising is the sharp increase in errors in this category in the large-γ columns. It is worth noting, too, that there is no penalty with re-

γ	0.01	0.1	0.25	0.5	0.75	1	2	4
Voicing	6	6	5	5	8	9	8	17
Manner	10	11	8	8	7	8	13	20
Place	8	7	3	3	2	5	18	12
Overall	18	16	12	11	15	18	28	34
Avg. Rank	1.44	1.38	1.26	1.25	1.28	1.41	1.88	2.31

Table 4: Results from experiment II. Classification errors and average rank are shown for various values of γ.

gard to the voicing decision for appropriate values of γ.

5 Conclusion

Audiovisual ASR offers increased accuracy and improved robustness compared to audio-only ASR. In order to fully exploit the potential of audiovisual ASR, it is necessary to develop effective strategies for integrating these two separate information sources.

The first class of strategies investigated involves integration as a final step. This class is the simplest to implement, and has been shown to yield significant performance improvement when compared to audio-only ASR systems; the variations explored here do not exhibit statistically significant differences from each other.

The second class of strategies uses a new HMM formulation wherein the classification is related to the joint probability of the visual and audio observations. Although the best performance of this method is approximately the same as that of the post-HMM integration method, there is reason to think that this is a pessimistic assessment, since the audio and visual data were not well synchronized.

Work is currently under way to construct a much larger, continuous speech database, so that these issues can be investigated with better statistical support.

Acknowledgement

This work was funded in part by NSF Grant IRI-9409851.

References

[1] P. L. Silsbee, *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition*, PhD thesis, University of Texas, 1993.

[2] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Proc. Intl. Conf. Acoust., Speech, Signal Processing*, pp. 557–560. IEEE, 1993.

[3] D. G. Stork, G. Wolff, and E. Levine, "Neural network lipreading system for improved speech recognition," in *Intl. Joint Conf. on Neural Networks*, pp. 285–295, 1992.

[4] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in B. Dodd and R. Campbell, eds., *Hearing by Eye: the Psychology of Lip-reading*, pp. 3–51. Lawrence Erlbaum Associates, London, 1987.

[5] J.-C. Junqua, H. Wakita, and H. Hermansky, "Evaluation and optimization of perceptually based ASR front end," *IEEE Trans. Speech and Audio Process.*, vol. 1, no. 1, pp. 39–48, Jan. 1993.

[6] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proc. Intl. Conf. Acoust., Speech, Signal Processing*, pp. 121–124. IEEE, 1992.

[7] P. L. Silsbee, "Motion in deformable templates," in *First IEEE Intl. Conf. on Image Process.* IEEE, Nov. 1994.

[8] R. Schwartz, S. Austin, F. Kubala, J. Makhoul, L. Nguyen, P. Placeway, and G. Zavaliagkos, "New uses for the N-best sentence hypotheses within the BYBLOS speech recognition system," in *Proc. Intl. Conf. Acoust., Speech, Signal Processing*, pp. 1–4, 1992.

[9] K. E. Finn and A. A. Montgomery, "Automatic optically-based recognition of speech," *Patt. Recogn. Lett.*, vol. 8, no. 3, pp. 159–164, 1988.

[10] K. W. Berger, *Speechreading: Principles and Methods*, National Education Press, Baltimore, 1972.