ROBUST AUDIOVISUAL INTEGRATION USING SEMICONTINUOUS HIDDEN MARKOV MODELS

Qin Su
Peter L. Silsbee

Dept. of Electrical and Computer Engineering
Old Dominion University

ABSTRACT

We describe an improved method of integrating audio and visual information in a HMM-based audiovisual ASR system. The method uses a modified semicontinuous HMM (SCHMM) for integration and recognition. Our results show substantial improvements over earlier integration methods at high noise levels.

Our integration method relies on the assumption that, as environmental conditions deviate from those under which training occurred, the underlying probability distributions will also change. We use phoneme based SCHMMs for classification of isolated words. The probability models underlying the standard SCHMM are Gaussian; thus, low probability estimates will tend to be associated with high confidences (small differences in the feature values cause large proportional differences in probabilities, when the values are in the tail of the distribution). Therefore, during classification, we replace each Gaussian with a scoring function which looks Gaussian near the mean of the distribution but has a heavier tail.

We report results comparing this method with an audio-only system and with previous integration methods. At high noise levels, the system with modified scoring functions shows a better than 50recognition does suffer when noise is low. Methods which can adjust the relative weight of the audio and visual information can still potentially outperform the new method, provided that a reliable way of choosing those weights can be found.

1. INTRODUCTION

1.1. Audiovisual Speech Recognition

Audiovisual methods of automatic speech recognition (ASR) have been widely studied as they offer improved robustness and accuracy, especially in the presence of noise. Traditional audio-based ASR systems perform reasonably well in controlled lab environments. In many environments, however, such as offices or outdoors, the recognition performance decreases drastically due to background noise. One way to increase robustness with respect to acoustic signal distortion is to consider the visual speech modality jointly with the auditory modality. Previous studies have shown that in both ASR and human speech perception, the audio and visual sensory modalities have different strengths and weaknesses, and in fact to a large extent they complement each other [7]. Visible speech is usually most informative for just those distinctions that are most ambiguous auditorily. For example, perceiving place of articulation, such as the difference between /b/ and /d/, is difficult via sound but relatively easy via sight. On the other hand, voicing information which is difficult to see visually is relatively easy to resolve via sound. Thus, visible speech is to a large degree not redundant with auditory speech.

The primary motivation for using visual information is to improve the robustness of the system with respect to environmental variations. Thus, a major goal is that an audiovisual system should perform at least as well as its audio subsystem does, over the entire range of conditions which might be encountered. This requirement implies that in situations where the audio subsystem performs accurately, the role of the visual information should be very limited, and as the audio subsystem loses accuracy, the role of the visual information should increase.

Since a system cannot "know" whether or not it is performing accurately, some measure of confidence must accompany the classification. A natural measure of confidence is the ratio of the highest score (probability estimate) to the nearest competing score. This confidence measure is easy to exploit such that when confidence for either subsystem (audio or visual) is high, then the decision of that subsystem carries a lot of weight, while if it is low, the other subsystem will have a substantial effect. Note that, in a phoneme-based HMM, this confidence and the associated decisions may be connected with individual states or time steps.

1.2. Integration of Audio and Visual Information

Several methods for integration of audio and visual sources have been proposed (e.g. [5, 4, 6, 9, 10, 11]). Robert-Ribes [4] has proposed a classification scheme for integra-

tion strategies. Two broad classes of strategy are "early" and "late" integration models. Early integration refers to strategies which combine evidence from different modalities prior to making any decisions, whereas late integration strategies perform some sort of independent single-modality scoring before combining evidence. Although there remains much to be discovered concerning this process, the evidence suggests that early integration strategies are the most successful.

2. INTEGRATION IN HMMs

Hidden Markov models provide a natural way of integrating audio and visual information with either an early or a late strategy. Early HMM integration methods are characterized, in the decoding process, by state-by-state estimation of observation probabilities based on audio and visual evidence. That is, in traversing the Viterbi lattice, both modalities are considered in determining the most likely branch leading to each node.

We now consider the problem of word recognition in hidden Markov models, using the Viterbi algorithm and its variants. These provide an efficient way of searching the very large space of possible state paths to find that path most likely to have generated the observation.

The Viterbi approach is particularly well suited for multimodal ASR. When recognition is based on the state path, it becomes possible to exploit subtle intermodal timing cues, since both signals are assumed to have been generated from a single state path [10].

Features derived from the audio and visual signals can become relatively more or less reliable due to insufficient training data and changes of environment; their relative influence on the recognition process should change accordingly. To allow for variability and changes in the relative reliability of the two information sources, some systems include one or more weighting parameters to control the relative influence of each signal [1, 7, 10, 12]. Typically, these weights are applied to either the output probability estimates at each time instant (in the case of early integration) or to the overall word scores (in the case of late integration). The weight associated with the audio signal increases as the acoustic signal-to-noise ratio increases. Other systems [2] operate without any such variable weighting parameter. Virtually all, it should be noted, use some form of product rule, based on an assumption of conditional independence between the audio and visual signals. That is, at some level, a score (usually a log probability) is computed according to

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

or a similar rule.

The use of variable weights introduces its own set of problems. For example, in order to know what values to assign to the weights, the system must be able to estimate the signal to noise ratio, or some other measure of signal quality. This is not easy to do, and if it is not done correctly, results can suffer. We seek a robust integration strategy which does not depend on variable weights, yet is suitable for use in a variety of environments.

If a HMM is trained with sufficient data which has been drawn from all potential operating environments, it should be able to make optimal decisions. A system operating in an unfamiliar environment is in essentially the same situation as a system which has simply been undertrained. In both cases, the probability models used to classify utterances poorly represent the "true" distribution which characterizes the operating environment. If discrete probability models are used, the result of this is simply that probability estimates are incorrect — the confidence associated with the estimate is generally unaffected. If continuous or semicontinuous models are used, there will be generally lower probability scores since the observations are likely to differ from those encountered during training.

It is important to understand the distinction between a decision made with *low confidence* and a decision made based on *low quality information*. Ideally, we would like these to be the same: when the information is poor, the decision should be relatively ambiguous. In particular, when the information from one modality is poor, that modality should have relatively little influence on the overall decision. However, in many cases, just the opposite is true. In the case described above — a model poorly trained for the environment in which it is operating — low probability estimates tend to be accompanied by inappropriately high confidence. The reason for this is that probabilities of independent observations are multiplied together. A single outlier in the observation sequence, which has a very low probability in all states, can exert more influence on the final decision than the high probability observations.

To see this, consider two univariate Gaussian pdfs $f_1(x)$ and $f_2(x)$, with means μ_1 and $\mu_2 = \mu_1 + \delta$, respectively, and identical variance σ^2. If an observed parameter value is x, then the overall effect of this one observation on the classification problem can be measured by the ratio $f_1(x)/f_2(x)$. If the ratio is close to 1, then the observation has little effect; if it is close to zero or very large, then the observation has a great effect. It is easily seen that the more distant x is from μ_1 and μ_2, the greater the effect it will have:

$$\frac{f_1(x)}{f_2(x)} \propto \exp\left(\frac{-\delta x}{\sigma^2}\right) .$$

If the variances are unequal, then the ratio grows as $\exp(x^2)$. While this would be perfectly appropriate if the trained distributions were accurate descriptions of the environment, it can have devastating consequences when the system is undertrained. And, as we stated above, we can consider any system to be undertrained when it encounters new operating conditions.

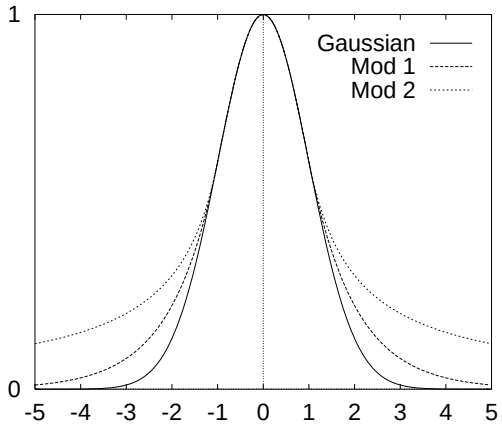

Figure 1: A Gaussian pdf and the two modified scoring functions described in the text. All are plotted to have a maximum value of 1. "Mod1" is the function with the tail modified to look Laplacian; "Mod2" (the slowest decaying function) has the tail modified to look like the reciprocal of the distance between the observation and the mean.

2.1. SCORING FUNCTIONS

Gaussian pdfs and Gaussian mixtures have been successfully used in many continuous and semi-continuous HMMs. In order to compensate for the state of undertraining which exists in variable environments, we have considered the use of modified scoring functions. These are essentially Gaussians which have had their tails replaced with functions that decay more slowly. Define

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \quad (1)$$

$$g(x) = \begin{cases} f(x), & |x-\mu| \le T \\ f(\mu+T)\exp\left[\frac{(T-|x-\mu|}{a})\right], & |x-\mu| \ge T \end{cases} \quad (2)$$

$$h(x) = \begin{cases} f(x), & |x-\mu| \le T \\ \left|\frac{Tf(\mu+T)}{x-\mu}\right|, & |x-\mu| \ge T \end{cases} \quad (3)$$

where $f(x)$ is of course a Gaussian density, $g(x)$ we refer to as the Gaussian/Laplacian function, and $h(x)$ we call the inverse distance function. Note that the modified functions are not true probability densities; nonetheless, they can be used in the scoring procedure as if they were ($h(x)$ cannot even be scaled to make a pdf). In order to avoid confusion, we will call these "scoring functions," whether they are true densities or not. Examples are shown in Fig. 1, scaled to have the same maximum value. The parameters T and a have been determined in our case from some pilot experiments; we have used values of T equal to 2σ and $a = T/2$, where σ is the standard deviation of the Gaussian obtained during training.

The important property that these two functions share is that the ratio $g_1(x)/g_2(x)$, for example, grows (or shrinks), much more slowly than does the equivalent ratio for the Gaussian density. The properties of the three functions as x grows large are summarized in Table 1.

Function	Equal variance	Unequal variance
Gaussian	$\exp(x)$	$\exp(x^2)$
Laplacian	constant	$\exp(x)$
Inverse distance	1	constant

Table 1: Behavior of the ratio of two scoring functions as x gets far from the center of the functions.

3. EXPERIMENTS

3.1. Experimental System and Database

The experimental data was described in detail in [7]. Ten audiovisual sequences were obtained from a single speaker for each of twenty-two consonants. Six tokens of each consonant were used for training, and four for testing. In other words, there were 22 different consonants, each repeated ten times, for a total of 220 tokens. 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 20 to 30 80×80 pixel frames, acquired at 30 frames per second. The audio data was sampled at 16,000 samples per second with 12 bits per sample. Differing amounts of white, Gaussian noise were added to the audio data to simulate different noisy environments. The audio processing uses a low-order RASTA-PLP speech processing subsystem [3]. The visual processor uses a simple deformable model [8] whose parameters after convergence represent several geometric features of the speaker's mouth.

3.2. Experiments and Results

Tests were run under the following conditions:

1. Semicontinuous HMM with only audio input.
2. Semicontinuous HMM with Gaussian scoring function, early integration.
3. Semicontinuous HMM with Gaussian scoring function, late integration, variable weight (the value of this parameter is determined automatically during training; see [10] for more details.)
4. Semicontinuous HMM with Gaussian/Laplacian scoring function, early integration.

Each system was tested with variable amounts of white, Gaussian noise.

Results of these experiments are summarized in Table 2.

The late integration model has the best performance, but this is because it makes use of a weight parameter determined *during training.* The weight parameter can, in principle, be adjusted during recognition to achieve similar results, but we do not have any reliable way to determine the appropriate value. The early integration methods outperform the audio system over most of the SNR range. However, the modified

	Condition			
SNR	1	2	3	4
0	8	12	33	15
5	16	17	33	26
10	23	32	44	40
15	39	44	56	55
20	63	64	69	64
25	73	77	85	74
30	89	80	91	77

Table 2: Summary of results (% correct) for the four experimental conditions listed in the text. Condition 1: Audio data only. Condition 2: Early integration using SCHMM and Gaussian scoring function. Condition 3: Late integration using weighting parameter determined automatically during training. Condition 4: Early integration using Gaussian/Laplacian scoring function.

scoring function appears to have an advantage only under noisier conditions. Obviously it is not acceptable for the visual information to degrade performance when the noise levels are low.

4. CONCLUSION

Audiovisual ASR methods have increasingly gained attention as they offer improved robustness and accuracy for ASR. We are continuing to explore ways of improving integration in audiovisual-based HMMs. In this paper, we have proposed the use of modified scoring functions based on Gaussian probability density functions, but with heavier tails to compensate for inadequate training of the Gaussian models. Note that this basic premise can be extended in many straightforward ways. In particular, we expect to implement similarly modified mixture densities. In addition, the choice of breakpoint between the Gaussian central portion of the function and the tail portion is admittedly somewhat arbitrary. It is possible that these breakpoints can be determined automatically as well (though we would expect only very minor effects from this).

The method we have described does not meet the critical criterion for audiovisual systems: that the combined system should perform at least as well as an audio system under all conditions which are likely to be encountered. This is unfortunate and can be seen as evidence that schemes with more adaptive ability will tend to be more successful. Nevertheless, the use of modified scoring functions has decreased the sensitivity of the system to environmental variations. A system combining adaptive weighting with robust scoring functions would be likely to outperform both.

5. ACKNOWLEDGEMENT

This work was supported in part by NSF grant IRI-9409851.

REFERENCES

1. Ali Adjoudani and Christian Benoît. On the integration of auditory and visual parameters in a HMM-based asr. In David Stork and Marcus Hennecke, editors, *NATO ASI: Speechreading by Humans and Machines*. Springer-Verlag, 1996.
2. Christoph Bregler, Stephen M. Omohundro, and Yochai Konig. A hybrid approach to bimodal speech recognition. In *Twenty-Eighth Annual Asilomar Conf. on Signals, Systems, and Computers*, pages 556–560, November 1994.
3. J.-C. Junqua, H. Wakita, and H. Hermansky. Evaluation and optimization of perceptually based ASR front end. *IEEE Transactions on Speech and Audio Processing*, 1(1):39–48, January 1993.
4. J. Robert-Ribes. A plausible and functional model for audiovisual integration. In David Stork and Marcus Hennecke, editors, *NATO ASI: Speechreading by Humans and Machines*. Springer-Verlag, 1996.
5. J. Robert-Ribes, Tahar Lallouache, Pierre Escudier, and Jean Luc Schwartz. Integrating audiotory and visual representations for audiovisual vowel recognition. In *Eurospeech*, pages 1753–1756, 1993.
6. J. Robert-Ribes, Jean Luc Schwartz, and Pierre Escudier. A comparison of models for fusion of the auditory and visual sensors in speech perception. *AI Review*, 1995.
7. P. L. Silsbee. *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition*. PhD thesis, University of Texas, 1993.
8. Peter L. Silsbee. Motion in deformable templates. In *First IEEE Intl. Conference on Image Processing*, volume 1, pages 323–327. IEEE, November 1994.
9. Peter L. Silsbee. Sensory integration in audiovisual automatic speech recognition. In *Twenty-Eighth Annual Asilomar Conf. on Signals, Systems, and Computers*, November 1994.
10. Peter L. Silsbee and Qin Su. Audiovisual sensory integration using hidden Markov models. In David Stork and Marcus Hennecke, editors, *NATO ASI: Speechreading by Humans and Machines*. Springer-Verlag, 1996.
11. Gregory J. Wolff, K. Venkatesh Prasad, and David G. Stork. Lipreading by neural networks: Visual preprocessing, learning, and sensory integration. In *Advances in Neural Information Processing Systems*, 1993.
12. Ben P. Yuhas, Moise H. Goldstein, Terence J. Sejnowski, and Robert E. Jenkins. Neural network models of sensory integration for improved vowel recognition. *Proc. IEEE*, 78(10):1658–1668, October 1990.