Sensor Fusion Weighting Measures in Audio-Visual Speech Recognition

Trent W. Lewis and David M.W. Powers

School of Informatics and Engineering
Flinders University,
GPO Box 2100, Adelaide, South Australia 5001,
Email: {trent.lewis|david.powers}@flinders.edu.au

Abstract

Audio-Visual Speech Recognition (AVSR) uses vision to enhance speech recognition but also introduces the problem of how to join (or fuse) these two signals together. Mainstream research achieves this using a weighted product of the output of the phoneme classifiers for both modalities. This paper analyses current weighting measures and compares them to several new measures proposed by the authors. Most importantly, when calculating the dispersion of the output there is a shift from analysing the variance to analysing the skewness of the distribution. Experiments in AVSR using neural networks raise questions of the utility of such measures with some intriguing results.

Keywords: Sensor Fusion, Speech Recognition, Neural Networks.

1 Introduction

The objective of Audio-Visual Speech Recognition (AVSR) is to enhance traditional speech recognition by incorporating a visual signal into the system. A simple way to achieve this is to combine both the acoustic and visual features into one large feature vector which is used for recognition. This technique is effective, given enough training data, but we can use knowledge from psychology and linguistics to conceive a more elegant system for combination. For example, it is known that visually perceivable speech gestures group into distinct classes of phonemes (known as visemes), and that these classes are complementary to speech sounds difficult to perceive in high acoustic noise (Walden et al., 1977). Sub-systems can thus be specialised for their modality and increase the overall system accuracy (Lewis and Powers, 2003). However, a one-to-many mapping exists between visemes and phonemes, so it may add another layer of complexity.

One of the most profound effects discovered in psycholinguistic research is the McGurk Effect (McGurk and MacDonald, 1976). If the audio of a person saying the sound "ba" is dubbed over the video of a person mouthing the sound "ga," the listener will perceive the sound "da." The brain has fused together the two competing signals. The effect is so strong that the researchers who discovered this at first thought the technicians had made a mistake. This result definitively showed that vision does have an influence over our perception of speech. The effect has also been extended to manipulate entire sentences (Massaro and Stork, 1998).

Research into machine AVSR has been very fruitful and systems have been developed showing very encouraging results (for a comprehensive review see Hennecke et al. (1996)). Although only minimal improvement is found under optimal conditions, improvements using a degraded acoustic signal have been large (Hennecke et al., 1996). For example, Meier et al. (1999) reported up to a 50% error reduction when vision is incorporated. However, a new problem also arises with AVSR, which is how to best combine the acoustic and visual signals without the result being worse than acoustic or visual recognition alone. This is referred to as *catastrophic fusion* (Movellan and Mineiro, 1998). This is a lively research area in AVSR and the effectiveness of different techniques, such as early, intermediate, and late fusion, are still being decided.

This paper briefly introduces the concept of sensor fusion with a more in depth look at current mainstream sensor fusion in the area of AVSR. Some of the more common techniques are then analysed and compared to several modifications to the standard algorithm.

2 Sensor Fusion

Information/Sensor/Data Fusion has had a long history, especially in the military domain. With the recent explosion in Data Mining, sensor fusion has been enjoying a renewed life with a focus on both expanding and refining data sets. Another area of sensor fusion that is also increasing in interest is the fusion of *ontological data*, and how this relates to the way in which the brain accomplishes this task given the enormous amount of fusion of sensory data that it performs.

2.1 Overview

Consider a sensor that has some unreliability associated with it and at times the output of this sensor is incorrect. When this error occurs is not known a priori. Thus, the sensor is basically useless, as we cannot determine when its readings are accurate. However, if we have another sensor with the same measuring ability (including its faults) then we can *more* reliably capture whatever it is we are sensing if we *fuse* the two sensor outputs to give one result. Increasing the number of sensors would increase the reliability providing the sensors/errors are (at least partly) independent. This type of fusion is known as *competitive fusion*, as the two sensors are competing to give the correct information, and works by using the *redundant* information contained in the overlapping sensors (Visser, 2001).

A key assumption in competitive fusion situations is that the noise contained in the output of the sensor is uncorrelated and independent from other sensors or classifiers (Kittler, 2000). Therefore when outputs are fused together the noise present will be cancelled out and the actual signal will be enhanced. However, if the noise present is correlated in some way, then the contribution of the noise to the output may actually be intensified.

A more interesting form of data fusion (and less affected by correlated noise) is known as *complementary fusion*. This is where one sensor has an incomplete or different view of the world whilst other sensors can complete the picture. Thus, each sensor contributes to give an overall picture of the world. This process can be slightly complicated by the fact that the sensing capabilities of each sensor may overlap and, more importantly, the sensor may be working with different representations, for example, a camera and a microphone.

Another issue that arises when different representations are involved is which representation to fuse in. One can fuse the data in each of the different representations, choose one representation as the base and convert all others to it, or choose an internal abstract representation that all sensor outputs are converted to. The latter two options are the preferred as this removes the conversion out of the fusion process, which is complicated enough.

Figure 1 is a schematic description of the difference between competitive and complementary fusion. On the left the two sensors are both attempting to identify a black square and thus a competitive fusion scheme would be used. Once that fusion has taken place the result is fused in a complementary fashion with the sensor on the right to complete the scene - a square and a triangle.

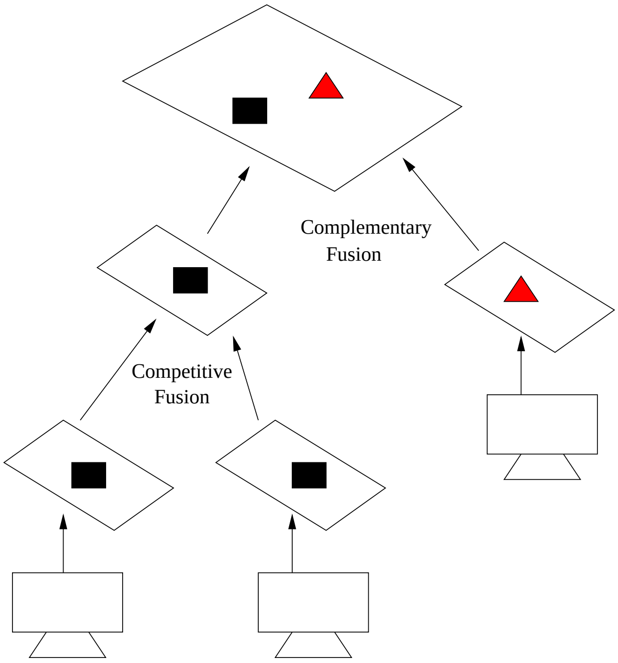

Figure 1: Example of competitive and complementary fusion

2.2 Sensor Fusion in AVSR

Initially this line of research investigated sensor fusion in the area of AVSR, however, sensor fusion is much broader than just AVSR and has applications in many domains. Moreover, this area of research is also known by many different names: classifier fusion, classifier combination, mixture of experts, committees of neural networks (NN), consensus aggregation, voting pool of classifiers, classifier ensembles, to name just a few (e.g., Kuncheva and Jain, 2000). Nonetheless, this section mainly focuses on sensor fusion in the domain of AVSR overviewing key aspects of different fusion systems.

Sensor fusion in AVSR broadly takes two different forms: early fusion (EF) and late fusion (LF). This differentiation is also called Feature/Decision fusion and Direct Identification/Separated Identification (DI/SI). Early Fusion is the case where features are extracted from the respective signals and then they are fused to create a combined feature vector that is used for recognition (Hennecke et al., 1996). Late fusion, on the other hand, is when feature vectors are extracted separately, classified separately, and then the results of the classifications (decision) are combined to give a final result. When following the DI model sensor fusion occurs automatically, and it is up to the recognition engine to decide upon the important features. This is the default approach if using ASR already.

Under the more sophisticated SI model, fusion becomes somewhat trickier. The simplest case is when the outputs of separate artificial NNs (ANNs) are fed into another ANN that effectively performs the fusion task. In the case of Hidden Markov Models (HMMs), the resulting log-likelihoods are combined in some way to produce a final estimate. The most common (and simplest) way to fuse the log-likelihoods is to combine them in such a way to maximise their cross-product. Late fusion (ie., SI) is an evolving area in AVSR and is a difficult issue to contend with because fusing the two signals can lead to *catastrophic fusion* (Movellan and Mineiro, 1998). This is when the accuracy of the fused outcome is less than the accuracy of both individual systems alone. Much work is underway for both HMMs and ANNs in trying to automatically bias one signal when conditions are adverse for the other (e.g. Adjoudani and Benoit, 1996; Massaro and Stork, 1998; Movellan and Mineiro, 1998).

There is still no consensus in the literature to when fusion should occur in the process. On theoretical grounds and the necessity of maintaining temporal relationships between the signals, many argue for early fusion (eg. Bregler et al., 1996; Basu and Ho, 1999). For example, Hennecke et al. (1996) state that late fusion is just a special case of early fusion and given the right conditions "... a system that uses early integration should perform at least as well as one that integrates at a later stage. (Hennecke et al., 1996, p. 338)." Indeed, if an inadequate set of sensor specific features are used, essential information can be thrown away in late fusion. Comparative empirical studies, however, have found that late fusion techniques are performing better than early fusion even with the loss of synchronisation (eg. Adjoudani and Benoit, 1996; Meier et al., 1999). The review that follows is mainly made up of research involving variants of late fusion as this technique has many more issues to overcome.

Potamiaonos and Potamianos (1999) use a multi-stream HMM in which the visual stream is just another parameter to the HMM. The emission probability of the HMM is equal to the product of the sum distributions of each stream, for example,

$$P(W|A,V) = \underset{W}{argmax}(P(W|A)^{\lambda_A} P(W|V)^{\lambda_V}) \quad (1)$$

These sum of distributions are augmented by a *stream exponent* λ. This exponent models the *reliability* of each stream and satisfies,

$$0 \leq \lambda_A, \lambda_V \leq 1, \quad and \quad \lambda_A + \lambda_V = 1 \quad (2)$$

The stream exponents are estimated using a generalised probabilistic descent algorithm. This appears to occur initially during training, but it is unclear as to whether the exponents are dynamically estimated during recognition. Thus in this system the late fusion is taking place via a weighted product of the contributions from the acoustic and visual channels. This is probably the most common approach to sensor fusion in this field and demonstrates that the AV system is superior to the acoustic or visual alone. Although the word accuracy by this system is high (90.5% for AV) the weights on each stream are determined a priori to test time (i.e. on the training set) and thus if the conditions change enough the weightings might not correctly reflect the reliability of each the signals.

Neti et al. (2001) and Glotin et al. (2001) have produced comparative studies of early, late with constant weighting, and late with *dynamic* weighting audio-visual fusion schemes. The dynamic technique was based on the degree of voicing present in the audio stream average over the entire utterance such that $0 \leq \lambda_A = degree\ of\ voicing \leq 1$ and $\lambda_V = 1 - \lambda_A$. Overall, the fusion system using the dynamic weights outperformed all others on a word recognition task in both clean and noisy acoustic conditions. Interestingly, in clean acoustic conditions some of the late fusion techniques were outperformed by the early fusion and in some cases even demonstrated catastrophic fusion.

Dynamically setting the weights based on the current utterance is a preferred method of fusion. This utterance based method, however, is somewhat lacking in its ability to generalise to other situations. For example, if there was a loud, brief sound in the background this might affect the overall average for the utterance and hence distort the weighting considerably. Calculating the median instead of a mean might correct the weights for the majority of the speech segment, but then at extra noisy sections performance would degrade. Dynamically determining the weights needs to occur at a lower level. Moreover, waiting until the end of the utterance to determine weights means that fusion can only take place after the *entire* utterance has been spoken.

Dupont and Leuttin (2000) tackle the problem of *continuous* speech recognition. In continuous speech recognition the system must deal with co-articulation and the fact that the utterance has no predetermined length. They claim that because of these factors waiting until the end of utterance to fuse is too time consuming for late fusion architectures and that fusion should occur during the utterance. Moreover, a list of the best hypotheses (the *N-best*) must be kept for each state until fusion occurs. Their speech recognition system consists of a multi-stream HMM with NN as HMM state probability estimators. This system uses *anchor points* to denote where individual streams must synchronise (fuse).

These anchors may occur on relevant phonological transition points, such as phonemes, syllable or words. Dupont and Leuttin (2000) only test anchor points at the HMM state and word level. Fusion is a weighted product of the segment likelihoods. These weights are determined by automatically estimating the acoustic signal to noise ratio (SNR), such that the higher the SNR, the higher the weight to the acoustic information. They mention that with a clean signal the addition of visual information did not increase accuracy. However, with a clean signal (high SNR) the weight was very high, and it might be that the visual system does not have the ability to influence the result given this weighting. Early fusion yielded inferior results compared to the different late fusion techniques. The most successful late fusion technique was with combination at the word level.

In their work, Adjoudani and Benoit (1996) strive for AV > A and AV > V over all testing conditions and explore several progressive models of fusion. The first, an early fusion method, fails in acoustically noisy conditions because it is dragged down by the inability of the system to capture the contribution of the visual parameters. The first late fusion technique is a simple maximisation of the product of the resulting probabilities across each output channel. In high SNR conditions the system is able to take advantage of the complementary information between the signals with AV outperforming both subsystems. In poor acoustic conditions, however, the system is once again not able to correctly attribute each subsystem.

To overcome the inadequacy of the combination so far, Adjoudani and Benoit (1996) introduced a *certainty factor* to differentially weight each subsystem. This weighting factor differs from previously discussed architectures as it is *not* solely based upon the level of acoustic noise within the signal. Rather, it is based upon the *dispersion* of the N-best hypotheses in each modality. Thus, large differences in probabilities equates to greater certainty, close probabilities to less certainty. This dispersion value is based upon the variance of the output classifier, as in

$$\sigma^2 = \frac{1}{N-1}\sum_{n=1}^{n}(R_n - \mu)^2, \tag{3}$$

where R_n is the n^{th} output of the classifier.

The first application of the certainty factor was a binary selection of either the acoustic or visual hypothesis based on which had the greatest certainty. This method satisfies the original criteria set by Adjoudani and Benoit (1996), however it can only ever choose between the classfication of the individual subsystems because of its binary nature. A weighted product version of the late fusion system based on a normalised dispersion certainty factor, as in

$$\lambda = \frac{\sigma_A}{\sigma_A + \sigma_V}, \tag{4}$$

combined the acoustic and visual systems synergistically over all noise levels and can choose a different class from either subsystem.

The *dispersion* idea used by Adjoudani and Benoit (1996) has been implemented by other researchers in various forms (e.g. Meier et al., 1999; Potamianos and Neti, 2000; Heckmann et al., 2001a). Using Gaussian mixture model (GMM) to classify phonemes, Potamianos and Neti (2000) use an N-best dispersion method that is framed as the difference between each pair of nth-best hypotheses, given by,

$$\frac{2}{N(N-1)}\sum_{n=1}^{N}\sum_{n'=n+1}^{N}(R_n - R_{n'}), \tag{5}$$

where N $\geq$ 2 and R_n is equal to the n^{th} best hypothesis. Interestingly, both Adjoudani and Benoit (1996) and Potamianos and Neti (2000) have found that an N-best of 4 has been the most successful. Potamianos and Neti (2000) also use a method called N-best likelihood ratio average in which the difference is only calculated against the best hypothesis, that is,

$$\frac{1}{N-1}\sum_{n=2}^{N}(R_1 - R_n), \tag{6}$$

where R is now sorted in descending order, such that this is the difference between the best hypothesis and the rest.

The best performing system here was the one using dispersion as a confidence measure with a phoneme accuracy of 55.19%. The ratio average achieved an accuracy of 55.05%. Both of these methods were significantly better than the baseline acoustic only system. Another confidence method based on the negative entropy of the stream was unable to achieve accuracy significantly better than the baseline.

Basu and Ho (1999) also used GMMs for recognition but only looked at early fusion. In comparison to Potamianos and Neti (2000), the accuracy of the system on the test data was consistently below 50%. Moreover, the combined feature vector provides little increase in accuracy. The value of this research however is that they also test the system on a *real-life* data set. That is, a data set not collected in a controlled environment and without specialised equipment. The performance on this data set drops dramatically with 33% for acoustic only and 9% for visual only. This clearly demonstrates that moving out of the experimental environment can severely affect even the "state-of-the-art" systems.

Heckmann et al. (2001a) use a hybrid ANN/HMM AVSR system with the NNs providing the a posteriori probabilities for the HMM which provide the phone and word models (language models). Heckmann et al. (2001a) argue for and use a late fusion method and use a weighting method they call *Geometric Weighting*. Detecting the most probable phoneme is found by a conditional probability that is augmented by the geometric weights. The value of the weight is based on another value c and they want c to reflect an estimate of the SNR of the acoustic signal. To achieve this they use a similar idea as dispersion by exploiting the distribution of the *a posteriori* probabilities at the output of the MLP, but based on the calculated entropy,

$$H = -\frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N}\hat{P}(H_{n,k}|\mathbf{x}_{A,k})log_2\hat{P}(H_{n,k}|\mathbf{x}_{A,k}), \tag{7}$$

where N is the number of phonemes and K is the number of frames. They created a mapping between c and H through an empirical analysis of the values (optimisation process). Results (Word Error Rates (WER)%) show a synergistic gain using this technique down to -6dB (high noise level) where it starts to perform worse than the visual. The automatic weighting performs similarly to manually setting c. They have also compared using entropy for setting c to using a Voicing Index and Dispersion methods, however, the entropy based c still gave the best results (Heckmann et al., 2001b).

Using a Multiple State-Time Delayed NN (MS-TDNN), Meier et al. (1999) utilise the flexibility of the NN to employ several different fusion methods for AVSR. They look at both the traditional early and late fusion but also fusion on the hidden layer of the NN. The early fusion technique included the standard concatenation and also the inclusion of an estimated SNR for the acoustic data. Late fusion is explored in two different architectures. The first is a weighted sum of the acoustic and visual systems. The weight was determined either by a piecewise-linear mapping to the SNR of the acoustic signal or by what they called "entropy weights". The calculation of entropy weights was not fully described in this paper (or previously for that matter, e.g. Meier et al., 1996), however, their description of the purpose of the weights, High Entropy = Even Spread = High Ambiguity = Low Accuracy, is reminiscent of the dispersion concept from Adjoudani and Benoit (1996). The entropy weights were further augmented by a bias b that "... pre-skews the weights to favour one of the modalities (Meier et al., 1999, p. 4)" This b value was set by hand to reflect quality of the acoustic data.

A more interesting and novel technique introduced by Meier et al. (1999) is the learning of the weights. They used another NN to combine both the acoustic and visual hypotheses with the output being the combined phoneme hypothesis. Theoretically, this technique should be able to at least match the performance of the other late fusion techniques as it can not only compute pair-wise comparison but also potentially make comparisons across the phoneme and viseme sets, thus taking advantage of the complementary information contained within the signal better than the simple weighted summation. In fact, best performance was with NN weight learning (except in high noise conditions). As would be expected from the bias b, entropy and SNR weighting performed similarly throughout. Early and hidden layer fusion combinations were, as others have found, poorer in performance.

Movellan and Mineiro (1998) compare standard Bayesian fusion technique (sum of log likelihoods) with what they call a robustified approach. They argue that most fusion system suffer from catastrophic fusion because they make implicit assumptions and degenerate quickly when those assumptions are broken and used outside its original context. The robustified approach makes these assumptions explicit by including extra parameters that represent the non-stationary properties of the environment. These parameters make up what is dubbed the *context model.* This approach works by not only maximising the probability with respect to the word but also to each context model (acoustic and visual). Movellan and Mineiro (1998) prove analytically that their approach is superior to the traditional as when the measurements yield data far from the model the traditional fusion system is heavily influenced by this subsystem. In contrast, the robustified approach limits the influence of signals far from a contextual model. Applied to AVSR using a HMM, this technique outperforms the classical in acoustic noise as well as with visual noise, an area not investigated by many researchers. In situations where normal fusion exhibits catastrophic fusion, the robustified fusion is no worse than acoustic or visual subsystems.

Not all of the research conducted follow the rigid late fusion architecture of weighted sum/product of hypotheses. For example, Verma et al. (1999) investigated audio-visual phone recognition using Gaussian mixture models with their second and third late fusion techniques being somewhat out of the ordinary. They look at three models of late fusion: 1) simple weighted sum, 2) weighted sum but V identifying only viseme and using an associated probability of the phoneme given the viseme, and 3) use both A and V to predict viseme (weighted sum, phase 1), then based on viseme class predict which phoneme class (weighted sum, phase 2). The sum of the weights was equal to 1 and was again adjusted manually. The recognition accuracies of the GMMs were well below that of systems combined with HMM. The third fusion technique (multi-phase) performed the best. However, this technique is not the most intuitive and a prime example of a system developed without linguistic knowledge. The very characteristic that is masked by noise in acoustic speech is the one that distinguishes the viseme classes (eg. /b/ from /d/, place of articulation), so that using hypotheses derived from the acoustic data in phase 1 could be more of a hindrance (although this isn't what is found in their experiments). Then in phase 2 they use V to distinguish *within* viseme classes! This is again very counterintuitive, given the definition of a viseme.

A more logical approach to fusion is presented by