Audio-visual Integration for Robust Speech Recognition Using Maximum Weighted Stream Posteriors

Rowan Seymour, Darryl Stewart, Ji Ming

School of Electronics, Electrical Engineering & Computer Science, Queen’s University Belfast, Belfast BT7 1NN, UK
r.seymour, dw.stewart, j.ming@qub.ac.uk

Abstract

In this paper, we demonstrate for the first time, the robustness of the Maximum Stream Posterior (MSP) method for audio-visual integration on a large speaker-independent speech recognition task in noisy conditions. Furthermore, we show that the method can be generalised and improved by using a softer weighting scheme to account for moderate noise conditions. We call this generalised method the Maximum Weighted Stream Posterior (MWSP) method. In addition, we carry out the first tests of the Posterior Union Model approach for audio-visual integration. All of the methods are compared in digit recognition tests involving various audio and video noise levels and conditions including tests where both modalities are affected by noise. We also introduce a novel form of noise called jitter which is used to simulate camera movement. The results verify that the MSP approach is robust and that its generalised form (MWSP) can lead to further improvements in moderate noise conditions.

Index Terms: Audio-visual integration, noise robustness, speech recognition.

1. Introduction

As early as 1984, Petajan [1] demonstrated that the addition of visual information can enable improved speech recognition accuracy over purely acoustic systems, as visual speech provides information which is not always present in the audio signal. Audio and video information can be integrated by feature fusion [2] or by decision fusion. Feature fusion assumes dependence between the audio and video streams and allows modeling of their correlation. Decision fusion assumes independence between the two streams and is performed by combining the results of separate classifiers for audio and video. Such classifiers are usually artificial neural networks [3, 4] or multistream hidden markov models [5, 6, 7, 8].

When both streams are uncorrupted, the audio features will usually provide better speech classification than the video features, so by using static fusion weights [5, 6] which favor the audio features, the combined recognition accuracy can be improved in comparison to a scheme where audio and video features are given equal weights. However, if the audio stream becomes less reliable due to noise, such a system will perform poorly, thus many systems have tried to estimate suitable weights dynamically. This has been done by estimating the audio signal-to-noise ratio (SNR) [3], and by examining the posterior probabilities [4, 7] or the class-conditional likelihoods of the individual streams at the frame level [8].

The maximum stream posterior (MSP) method which was introduced in [9], differs from these in that it automatically estimates the optimal weights for the combination, assuming a binary weighting scheme, which either uses or discards a stream depending on its reliability measured in terms of the combined stream posterior probability. This MSP method was previously tested positively on a small speaker-dependent database involving both audio (e.g., background noise) and video (e.g., compression, blurring) corruptions. In this paper we demonstrate that the MSP method works well in speaker independent tests using a much larger database. We also introduce a new form of video corruption called jitter which we use to simulate camera movement. Furthermore we develop a generalisation of the MSP method which involves relaxing the hard weighting scheme to a softer more flexible weighting scheme to allow a partially corrupted but still useful stream to be included, based on a maximum posterior probability criterion. The intention is that the weights will reflect the relative reliability of the individual streams in an optimal way. We refer to this new method as maximum weighted stream posterior (MWSP).

Speaker-independent digit recognition experiments were conducted on the large XM2VTS database [10], to assess the performance of the MSP and MWSP methods in various noisy conditions. We also compare these methods with the posterior union model [11], which is an alternative way of implementing soft weighting in noisy conditions. Importantly, all of these methods assume no prior knowledge of the noise type or level on either stream.

2. Audio-Visual Integration

2.1. Maximum Stream Posterior (MSP)

An ideal audio-visual integration approach should satisfy two criteria. Firstly, to outperform either stream on its own for low levels of corruption in either stream, and secondly, in cases where one stream is highly corrupted, to perform similarly to the remaining clean stream. The MSP method [9] seeks to meet these criteria by selecting either the audio stream $o^A$, the video stream $o^V$ or the combination $o^{AV}$ by maximizing the posterior probability of the speech state for each frame. Let $P(s|o)$ denote the optimal posterior probability of state $s$ given frame $o$, where $o$ could be $o^A$, $o^V$ or the combination $o^{AV}$. The MSP method can be expressed as

$$P(s|o) = \max\{P(s|o^A), P(s|o^V), P(s|o^{AV})\}$$  \hspace{1cm} (1)

where $P(s|o^A)$, $P(s|o^V)$ and $P(s|o^{AV})$ are the posterior probabilities of state $s$ given the individual streams or the combination. Using Bayes’ theorem they can be written as:

$$P(s|o^A) = \frac{P(o^A|s)P(s)}{\sum_{o'} P(o'|s)P(s)}$$  \hspace{1cm} (2)
\[ P(s|o^V) = \frac{p(o^V|s)P(s)}{\sum_{s'} p(o^V|s')P(s')} \tag{3} \]
\[ P(s|o^AV) = \frac{p(o^A|s)p(o^V|s)P(s)}{\sum_{s'} p(o^A|s')p(o^V|s')P(s')} \tag{4} \]

where \( p(o^A|s) \) and \( p(o^V|s) \) are the likelihood functions of \( o^A \) and \( o^V \) and independence is assumed between them, \( P(s) \) is the prior probability of state \( s \), and the summation in the denominators for \( s \) is over all possible states within the search beam.

The optimal posterior \( P(s|o) \) will be incorporated into an HMM as an approximation of the state-based emission probability.

If it is assumed that the least corrupted stream will produce the maximum ratio of likelihoods between the correct state and the remaining incorrect states, then selecting the maximum of the posteriors \( P(s|o^A) \), \( P(s|o^V) \) and \( P(s|o^{AV}) \) is likely to obtain the least corrupt stream. This can be shown by rewriting the posterior probabilities in a form of likelihood ratios between the states. For example, in the case of \( o^A \) we can rewrite Eq. (2) as:

\[ P(s|o^A) = \frac{P(s)}{\sum_{s' \neq s} p(o^A|s')P(s')} \tag{5} \]

For correct state \( s \) the likelihood ratio in the denominator:
\[ p(o^A|s)/\sum_{s' \neq s} p(o^A|s') \]
and hence the posterior probability \( P(s|o^A) \), is likely to be maximized when \( o^A \) is the least corrupt stream.

Therefore the MSP method (Eq. (1)) represents a method for choosing the optimal feature stream for recognition without assuming prior information about the corruption.

2.2. Maximum Weighted Stream Posterior (MWSP)

A potential weakness of the above MSP method is that a stream either contributes equally to the final posterior probabilities, or is ignored completely. It is intuitive that if one stream is clean and one has moderate corruption, the latter can still contribute useful information, but that less confidence should be placed on it, i.e., a weighting against it. The MWSP method seeks to find a softer and optimal weighting for the combination of the two streams by examining a set of weightings which covers the full range of relative stream confidences. This range includes equal confidence (equivalent to Eq. (4) of the MSP), and absolute bias toward either stream (equivalent to Eq. (2) and (3) of the MSP), and thus the MSP can be considered a special case of the MWSP. The MWSP is described by the following general formula which gives the posterior probability of state \( s \) for a given weighting \( w \),

\[ P_w(s|o) = \frac{p(o^A|s)^wp(o^V|s)^{1-w}P(s)}{\sum_{s'} p(o^A|s')^wp(o^V|s')^{1-w}P(s')} \tag{6} \]

where it is assumed that \( w \in [0, 1] \). The MSP method described in Section 2.1 can be obtained by setting \( w = 1 \) (Eq. (2)), or \( w = 0 \) (Eq. (3)), or \( w = 0.5 \) (Eq. (4), approximately). The optimal posterior probability used for recognition is the maximum across all of the weightings, i.e.,

\[ P(s|o) = \max_w P_w(s|o) \tag{7} \]

2.3. Posterior Union Model (PUM)

The PUM is an alternative way of implementing softer weighting, based on the probability theory for the union of random events. It has been shown to increase the accuracy of acoustic speech recognition in the presence of partial frequency corruption [11]. For this application of audio-visual integration a model with two possible combinations of streams is used. These correspond to the conjunction and disjunction of the two streams, i.e.,

\[ P_w(s|o) = \frac{p(o^A|s)p(o^V|s)P(s)}{\sum_{s'} p(o^A|s')p(o^V|s')P(s')} \tag{8} \]
\[ P_v(s|o) = \frac{p(o^A|s) + p(o^V|s)P(s)}{\sum_{s'} (p(o^A|s') + p(o^V|s'))P(s')} \tag{9} \]

The maximum of \( P_w(s|o) \) and \( P_v(s|o) \) is selected as the state-based emission probability within a HMM.

3. Database and Acoustic Modeling

In our previous work [9], the MSP method was tested on a small speaker-dependent data set. In this paper we used the considerably larger XM2VTS database [10] which allowed all of the aforementioned methods to be tested and compared on speaker-independent digit recognition tasks. The database contains 295 speakers, roughly balanced between genders. Each speaker was recorded saying all ten digits four times in four different sessions in a quiet environment. The data was divided into 200 speakers for training and 95 speakers for testing. Thus, there are 3200 training occurrences of each digit and the test data includes 15200 test tokens.

The database was supplied with some lip tracking results, using the colour based approach described by Ramos [12]. These were used to localize the mouth region of interest (ROI) in each video frame, eliminating the need for mouth tracking. Previously in [9], a geometric based visual feature extraction method was used, which involved fitting an active contour based model of the lips to the video frame, and taking measurements from that to form the visual feature set. Our preliminary experiments using the XM2VTS database however, showed that image transform (or appearance) based features could provide both greater accuracy and significantly faster extraction times. This is consistent with the findings of other researchers [13, 14]. Thus a DCT (type II) image transformation was implemented. This transform has been shown to outperform other image transforms such as the DWT [13].

For each frame of video, the mouth ROI was cropped and subsampled to 16 × 16 before applying the DCT as shown in Figure 1. The transformation results in a 16 × 16 array of coefficients for each frame of video. The visual features were selected from these coefficients using a triangle mask (see Figure 1) as this reflects how the coefficients are arranged with the lowest frequencies at the origin. Finally, the visual feature stream \( o^V \) was formed by interpolating these features to 100 Hz to match the sampling rate of the audio feature stream \( o^A \). Cubic splines were used to interpolate the features, which enabled easy calculation of dynamic features, i.e., first and second derivatives (\( \Delta \) and \( \Delta \Delta \) respectively). Experimentation with different mask sizes and combinations of static and dynamic features showed that the optimum DCT feature selection was 15 static and 15 dynamic features. The DCT features were normalized using mean subtraction to account for inter-speaker and session variability.

The audio features were extracted using a 30 channel mel-scaled filter bank from which 10 mel-frequency cepstral coefficients (MFCCs) plus energy were taken. The corresponding \( \Delta \) and \( \Delta \Delta \) dynamic features were also calculated, to form the audio feature vectors. Each digit was modeled using an HMM with 10 states (no state skipping), with 4 mixtures per state.
4. Experimental Results

We experimented with the use of different quantization levels for $w$ from 4 up to 10 within the range $[0, 1]$, and they were all found to produce better results than the binary weighted MSP. In the following experiments the MWSP method was implemented with 8 equally spaced values for $w$ within $[0, 1]$. As mentioned earlier, $w = 0$ is complete bias toward the video stream, and $w = 1$ is complete bias toward the audio stream.

4.1. Audio Corruption

To examine the effect of corruption in the audio stream, additive white noise was added to the test data at different SNR levels (-20dB to 30dB). Recognition accuracy results (in WER) are shown in Figure 2 for audio and video separately, as well as the three integration methods, MSP, MWSP and PUM. It can be seen that the audio features alone give very good accuracy at the highest SNR level, but the accuracy level drops rapidly as the SNR level decreases. The video features alone, unaffected by the audio stream corruption, provide a constant WER of 14.82%. At all but the 2 lowest SNR levels, the MSP, MWSP and PUM methods outperform either stream on its own, with the MWSP method performing slightly better than the others in most conditions.

These results indicate that the MWSP is selecting stream weights which more accurately reflect the relative reliabilities of the two streams than the MSP method’s hard weighting scheme and hence it allows both streams to be more effectively combined in moderate noise conditions.

To investigate this further, we recorded the average stream weights which the MWSP method selected for the most likely state at each frame in our tests at different levels of audio SNR. The results are shown in Figure 3. It can be seen that as the audio SNR increases, so too does the average weight which is applied to the audio stream. This shows that the MWSP method can automatically adapt to the relative reliability of the streams in a smooth way.

4.2. Video Corruption

To examine the effect of corruption in the video stream, two different types of noise were used: blurring and jitter. Blurring is performed by applying a Gaussian filter to each video frame and represents real world situations where the video camera loses focus on the speaker. Jitter is a novel form of corruption we developed to simulate situations where either the camera or the speaker are moving, preventing smooth tracking of the mouth. Jitter is generated by translating and rotating the mouth ROI by a random amount. Figure 4 shows examples of the effect of blurring and jitter corruption on the video samples used in the recognition.

Combining the corrupted video features with clean audio features provided almost no discrimination between the MSP and MWSP methods, as both methods performed very closely to the audio-only system, which achieved 1.38% WER. Further experiments were conducted with corruptions in both the video features (as above) and the audio features (at 15 dB SNR).
Figures 5 and 6 show the recognition accuracy of the different methods with blurring and jitter respectively. It can be seen that the MSP and MWSP methods are robust to these forms of corruption. It was also found that the MWSP method again provided a small improvement upon MSP for almost all levels of video corruption (although this is not highly visible in the Figures due to the scale required). These results again show that the softer weights selected by MWSP can be beneficial compared to the hard weights selected by MSP. The PUM, displays robustness in these tests but does not perform as well as the other methods at any level of video noise.

5. Conclusions

In this paper we have demonstrated the robustness of the Maximum Stream Posterior (MSP) approach to AV integration for speaker independent speech recognition on a large digit database. Furthermore we have presented a more general form of MSP we call the Maximum Weighted Stream Posterior (MWSP) method. This approach allows a softer more flexible stream weighting scheme to be applied than the MSP approach which can be beneficial in moderate noise conditions. We have demonstrated that this approach automatically adapts to the stream noise levels and can provide improved recognition accuracies. Both these methods have been compared to the Posterior Union Model method in tests involving corruption in either/both stream and are shown to be more robust for audio-visual integration for multimodal speech recognition in noisy conditions.

6. References