

NOISE-BASED AUDIO-VISUAL FUSION FOR ROBUST SPEECH RECOGNITION

E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy

Department of Electrical and Computer Engineering
Clemson University
Clemson, SC 29634, USA
{epatter, sabrig, ztufekci, jgowdy}@eng.clemson.edu

ABSTRACT

A major goal of current speech recognition research is to improve the robustness of recognition systems used in noisy environments. Recent strides in computing technology have allowed consideration of systems that use visual information to augment the decision capability of the recognizer, allowing superior performance in these difficult environments. A crucial area of research in audio-visual speech recognition is how to combine the separate modes of information.

Late integration, an approach whereby separate audio-based and video-based decisions are made and then combined “late” in the process, has emerged as one of the simplest yet most effective techniques. Research has suggested that the fusion method for this technique (and similar methods such as multi-stream HMMs) is affected somewhat by the level of interfering audio noise. This paper further defines the relationship between data fusion in the presence of audio noise and demonstrates that optimal data fusion can only be performed if both the noise level and type are considered.

1. INTRODUCTION

Speech recognition research has made tremendous progress in the last few decades, resulting in very high accuracy rates but only under certain constraints. Typically, a clean speech environment is needed. For most applications, though, a speaker does not have this luxury. There is usually some type of corrupting background noise in vehicles, homes, shopping centers, or other locations where speech recognition could serve in useful applications.

The most significant problem caused by the presence of noise is the loss of information necessary for speech recognition decisions. Features used for audio speech recognition fail to distinguish elements of speech, particularly when the energy of the corrupting noise approaches or surpasses that of the speech. There are techniques for helping to remove noise or its effects but these typically remove important speech information as well. Other techniques may increase the robustness of features, but results are often well below those obtained in a clean environment. Matched training helps overcome many of these problems by training the system on speech corrupted by the same noise present in the environment. This is often not practical, though, since a user’s environment is very likely to change in many desired applications. An ideal system should be able to be trained in a clean environment and used in many different types and levels of noises while still maintaining acceptable recognition rates.

In order to improve system performance in difficult environments, it is necessary to provide as much information as possible to the recognizer despite noise. An approach that has shown posi-

tive results for overcoming the negative effects of noise is the use of both audio and video features for speech recognition. The addition of video information that is not corrupted by background noise allows for improved overall results. It has also been shown that lipreading plays an important role in human speech recognition. A common example is the McGurk effect, where a spoken phrase such as “my bab pop me poo brive” overlaid on a video of a speaker saying “my gag kok me koo grive” is actually understood as “My dad taught me to drive.” Without the video, though, the audio is considered unintelligible nonsense[1]. This demonstrates the power that lipreading has to influence human audio perception. Other results have also shown that people benefit from lipreading while attempting to understand speech in noise[2]. In addition, confusion trees have been created for audio and visual speech to show that different information is provided by each[3]. This ability to supplement audio-based features with additional information from video should allow significant improvements in speech recognition, particularly in the presence of audio noise, where the video is not corrupted. Petajan and others have developed audio-visual systems and demonstrated good recognition results[4][5][6]. Typically, the addition of video information does improve the recognition rates.

One focus of research in audio-visual speech recognition is the combination of the separate streams of information. There are three main approaches to audio-visual fusion, based on when the combination occurs. These techniques have been compared against each other, indicating that late integration is possibly the most effective, despite its simple approach[5]. In late integration, separate audio and video decisions are made and combined as the last step in the system. The combination may be performed by a simple fuzzy *and* rule, related to the Bayesian decision rule, which weights the audio and video decisions for a joint decision[1][7]. This is also consistent with the Fuzzy Logical Model of Perception (FLMP) that suggests that humans have separate evaluation, integration, and decision of linguistic choices from multiple sources of information[8]. Intuition would suggest that the video information should become more important as the audio information is corrupted. This can be viewed as a confidence measure of the video and audio features based on the noise level (SNR).

This paper demonstrates that for an unmatched speech recognition system there is a direct, nearly linear relationship between the ratio used in the decision rule for late integration and the noise type and level. Tests with our audio-visual speech recognition system were performed adding each of the noises from the NOISEX database at various SNRs to the test speech and not the training speech. The optimal values for audio-visual fusion were found and studied for a correlation to the type and energy of noise applied to

the speech. The following section will present our audio-visual recognition system and some of its benefits. Section 3 will discuss some of the issues of audio-visual fusion, late integration, and the fuzzy multiplicative rule used for combining audio and visual decisions. Following that, section 4 will present information on our system and testing, as well as results from our studies. Finally, section 5 will summarize our results and conclusions, mainly that the noise type and level directly affect the combination of video and audio information in an unmatched system used in noisy conditions by acting as a measure of reliability of the audio decision. This could lead to a dynamic decision method for the fusion ratio in audio-visual speech recognition systems in order to provide superior noise robustness.

2. AUDIO-VISUAL RECOGNITION SYSTEM

The audio-visual recognition system used for this research is a combination of separate audio and video HMM-based speech recognizers. Each system produces log probability scores for each possible word, and these are combined after both systems have finished processing. The decision rule used for data fusion is a fuzzy *and* rule common to late integration in audio-visual speech recognition. Late integration is discussed further in section 3. The combined decision will ideally yield more accurate results than either individual system.

The video subsystem uses color video sampled at thirty frames per second. The lips are located in the image by dividing the red plane by the green plane and thresholding before edge-detection[6]. After the lips are extracted, video features are found based on the oral cavity (lip width, height, width-to-height ratio, and open mouth area) and affine-invariant Fourier descriptors. The four oral cavity parameters and twelve Fourier descriptor coefficients are passed to the HMM for recognition. The use of Fourier descriptors for audio-visual speech recognition is a novel approach that appears to yield good results that are robust to movement of the speaker[9].

The audio subsystem uses audio sampled at 16 kHz. Sixteen mel-frequency discrete wavelet coefficients are used for audio features[10]. The audio subsystem is trained on clean speech, and then noise is added artificially from the NOISEX database. Noise is added to create test speech samples with varying SNRs. The system is then tested on these samples and clean speech. The separate N-best choices from each subsystem are used in late integration to produce the final decision.

3. LATE INTEGRATION BASED ON SNR

The goal of audio-visual data fusion in speech recognition is to yield recognition rates that are higher than possible with either individual subsystem. There are three major approaches for when audio-visual fusion occurs. One technique is direct identification. It works by passing both streams of information to a bimodal classifier, so the fusion actually occurs during classification. Another technique is early integration (or data-to-data); here, fusion of the audio and video information occurs before classification. There are multiple models for this technique, such as dominant recoding and motor recoding. Dominant recoding supposes that audio dominates video in speech perception and recodes the video information to some form of audio representation. Another model for early integration is motor recoding, which transforms both audio and video information to a common space related to speech gestures[5]. Finally, there is late (or separate) integration, where the data fusion occurs after decisions are made. This allows one of

the simplest, yet most effective, fusion methods. Some have presupposed that some information from the characteristics of video and audio has been lost at this point, but results for late integration seem to equal or surpass other techniques[5]. There is little doubt that any feature-based recognizer will lose some information, but late integration performs very well and is also simpler to implement and control. Similar results are obtained using multi-stream HMMs, moving the data fusion earlier in the process, to aid audio-video synchronization in continuous speech recognition systems.

The actual fusion in late integration is usually performed by a simple fuzzy logic multiplicative rule related to the Bayesian decision rule[7][1]. For example, if A_i is an indication that word i is "heard" by the audio subsystem and V_i is an indication that word i is "seen" by the video subsystem, the combined "truth" of word i being perceived is $t(P_i)$ as seen in equation 1.

$$t(P_i) = t(A_i)t(V_i) \quad (1)$$

Truth scores are treated similar to probabilities, ranging from 0 to 1, and assumed independent in a multiplicative *and* rule. Because HMM-based recognizers usually produce log probability scores, though, the relationship is actually expressed in terms of addition in order to remain in log domain. (Resulting probabilities are quite low, otherwise.) This is expressed in equation 2 where S_{ai} and S_{vi} are the log probability scores from the recognizer and may be viewed as $\log(t(A_i))$ and $\log(t(V_i))$.

$$\log(t(P_i)) = S_{ai} + S_{vi} \quad (2)$$

Rather than strict independence in the fuzzy logic multiplicative rule, a quality score will be given each information stream, called λ_a and λ_v . Typically, the value λ_v is actually expressed as $(1-\lambda_a)$ in keeping with the fuzzy logic "truth" system. Hence, $\hat{t} = t(A_i)^{\lambda_a}$ and $\hat{t} = t(V_i)^{\lambda_v}$, which yields equation 3 and the equivalent equation 4 where $\lambda = \lambda_a$.

$$\log(\hat{t}(P_i)) = \lambda \log(t(A_i)) + (1 - \lambda) \log(t(V_i)) \quad (3)$$

$$S_i = \lambda S_{ai} + (1 - \lambda) S_{vi} \quad (4)$$

As mentioned in [7], a value of $\lambda = 0.5$ may not be "equal" weighting of video and audio information streams in a system of this type. In our system, video log scores are small positive values with 100 points or so between word scores, but the audio log scores are large-magnitude negative numbers with 1000 points or so between word scores. This is due to the HMM-based recognizer implementations.

Several researchers have looked at the SNR of accompanying audio noise but not at the relationship of the type of noise to the effects on data fusion. Silsbee and Bovik suggested in their work that they found no relationship between noise levels and optimal choice of λ . [7] They suggested that a single value for λ worked equally well for all noise levels. However, they also tested on the same conditions with which they trained, so their results are based on a system matched to its test environment which does not allow the independent information from the video stream to aid the overall results as much. Others have noted that there is a direct relationship between the SNR of accompanying noise and the ratio used for data fusion but did not investigate varying types of noises [11] [12].

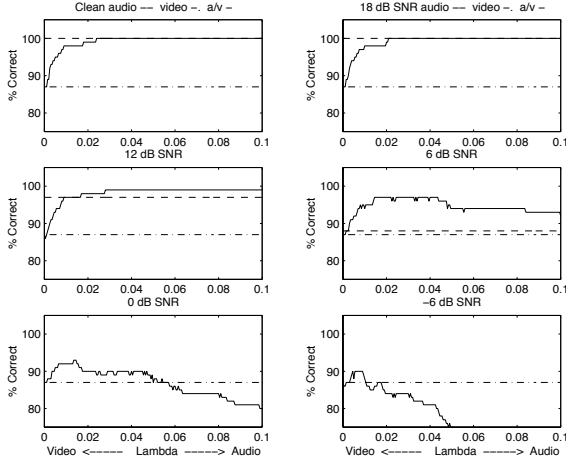


Fig. 1. Audio-visual recognition rate versus λ .

4. RESULTS AND OBSERVATIONS

The system described in section 2 was used to test for a relationship between λ and the SNR. Each noise from the NOISEX database (speech, Lynx helicopter, operations room, machine gun, STITEL, F-16 jet, factory, and car noise) was added to the clean speech at signal-to-noise ratios of 18dB, 12dB, 6dB, 0dB, and -6dB. The training and test database consisted of speaker-dependent samples recorded in our lab, using color video and 16 kHz audio. Seventeen sets of a ten-word vocabulary (down, forward, no, off, on, radio, rewind, tape, up, yes) such as to control a car stereo (and similar to those used in [6]) were recorded for both video and audio. The training was performed on sixteen sets and tested on one. Training and testing sets were rotated seventeen times to improve testing accuracy without recording many more individual recordings. All audio-HMM training was performed on clean sets only. Both audio and video subsystems used sixteen coefficients as described in section 2. Each HMM subsystem was created in HTK using eight-state, left-to-right word models.

The audio and video subsystems were tested separately and then jointly based on all possible λ values with an incremental resolution of 0.0005 for λ testing. The audio was tested on clean speech and then on each SNR for each noise. The video remains the same for each case, as it is unaffected by the audio noise. The audio subsystem obtained 100% accuracy for the clean speech case, and the video subsystem obtained 87% accuracy. Several plots follow which help establish a relationship between the SNR and an optimal value for λ for each noise type.

Figure 1 shows the general trend between recognition rates and λ . The graph is plotted with a small-valued axis, so that the trend concerning the recognition rates and λ may be seen. The dashed line is the audio performance, dot-dash the video performance, and solid line the audio-visual performance. As the noise increases between each plot, the dashed audio line drops, eventually below that of the video which maintains the constant 87%. As the noise level increases (SNR decreases), the range of best recognition performance becomes both smaller and farther to the left on the graph. This indicates that a smaller value for λ is necessary to obtain the maximum recognition rate when the noise level is increased. With a smaller value for λ , the system will weight the

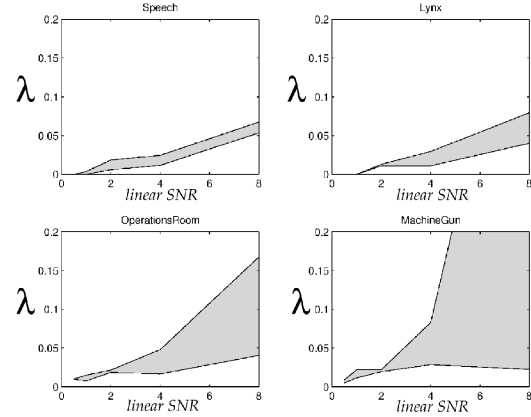


Fig. 2. Optimal values (shaded region) of λ for each noise versus linear SNR.

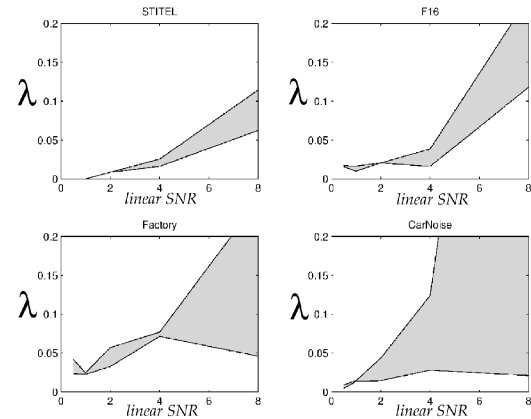


Fig. 3. Optimal values (shaded region) of λ for each noise versus linear SNR.

video score more to compensate for the increased noise.

Figures 2 and 3 show the optimal ranges for λ for all types of noise from the NOISEX database by plotting λ versus the linear signal-to-noise ratio. Any values within the approximate projected range should give the highest audio-visual recognition rates (Only SNRs of 18dB, 12dB, 6dB, 0dB, and -6dB or 8:1, 4:1, 2:1, 1:1, 1:2 have been determined so far, not a continuous range.) It can be seen that there is a direct, nearly linear relationship between the linear SNR and optimal λ value. In fact, figure 4 demonstrates that a simple line equation can provide optimal values of λ in most cases. It is important to note, though, that the range of optimal values of λ varies for each noise type. Some are more narrow than others, and some have steeper slopes. This would indicate that a dynamic estimation of both the SNR and some classification of the noise would result in improved results over static methods.

Based on these graphs, some observations may be made about the relationship between the multiplicative decision rule and the SNR. There does appear to be a direct correlation that in most cases may be solved optimally by a simple line equation. The regions for optimal values, however, vary for each noise type.

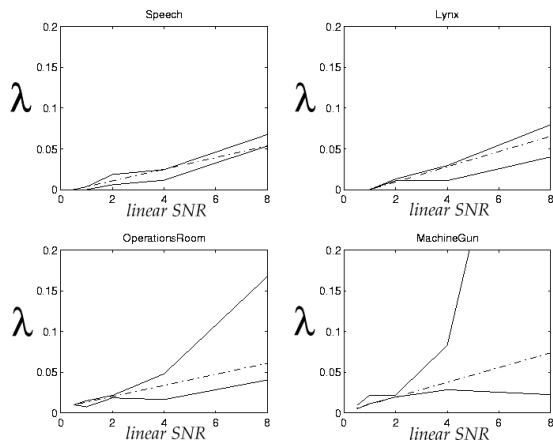


Fig. 4. Linear solution for optimal values of λ .

	audio %	video %	av %
clean	100	87	100
18 db	96.5	87	99.3
12 db	80.9	87	96.9
6 db	52.6	87	92.9
0 db	29.1	87	89.8
-6 db	21.0	87	88.4

Table 1. Recognition Rates Averaged Over Noises

The slope appears to consistently change past 12dB (4:1) to 18dB (8:1); in several cases, the optimal region widens significantly here as well, implying that either video or audio decisions will work well with high SNRs for the particular noise. For instance, with car noise the system would do quite well “leaning” on either subsystem during cleaner environments versus speech noise where it needs to rely more on the video subsystem for optimal results. The increased slope in the solution line and the whole region of optimal values of λ , though, coincides with placing more importance on the audio information in cleaner environments. The slope of the solution line and also the whole region for optimal values of λ also appeared to vary somewhat based on each type of noise. (Implementing a noise detector that changes the dependence of λ on the SNR based on each noise could possibly give good results in a practical system.) For any of these noises, though, using an optimal λ from the range based on the noise type and SNR results in better combined recognition rates than a constant λ . (Using a fixed decision rule or late-integration based solely on the SNR did not yield optimal results in our studies.) This may be seen by observing the regions in the chart and also noted by referring to the recognition rates in table 4. These rates include audio, video, and audio-video results averaged over all types of noises tested and demonstrate that basing λ on the SNR provides combined results that exceed either subsystem in all noise levels. (In fact in testing, it was noted that, even in significant noise, as long as the audio rate was at least 20% correct, it was still providing information to allow the audio-visual system to exceed the constant 87% accuracy of the video subsystem.)

5. SUMMARY AND CONCLUSIONS

This paper investigates the affects of noise type and level on a speech recognition system that uses late integration to combine separate audio and video decisions. The decision rule for integration is a fuzzy multiplicative *and* rule, weighting the audio score by λ and the video score by $(1-\lambda)$. Results are shown that demonstrate that the optimal value for λ is related to both the noise type and level of the interfering noise.

By choosing the value λ based on optimal regions for each noise and SNR, the system is able to obtain optimal recognition rates which usually surpass those of either subsystem but never fall below the rate of the strongest subsystem. The SNR acts as a “reliability” measure for the audio and allows the system to compensate for noisy audio by more heavily weighting the video decision to prevent the combined recognition rate from dropping below the video rate. These results indicate that the optimal value for λ does vary and would be best chosen using a dynamic algorithm based on both noise classification and SNR estimation.

6. REFERENCES

- [1] D. W. Massaro and D. G. Stork, “Speech recognition and sensory integration,” *American Scientist*, vol. 86, 1998.
- [2] Q. Summerfield, “Lipreading and audio-visual speech perception,” *Phil. Trans. R. Soc.*, vol. 335, 1992.
- [3] S. Nishida, “Speech recognition enhancement by lip information,” 1986.
- [4] E. Petajan, B. Bischoff, D. Bodoff, and N. Brooke, “An improved automatic lipreading system to enhance speech recognition,” in *ACM SIGGHI*, pp. 19-25, 1988.
- [5] P. Teissier, J. Robert-Ribes, J. Schwartz, and A. Guérin-Dugué, “Comparing models for audiovisual fusion in a noisy-vowel recognition task,” *IEEE Transactions on Speech, and Audio Processing*, vol. 7, no. 6, 1999.
- [6] G. I. Chiou and J. Hwang, “Lipreading from color video,” *IEEE Transactions on Image Processing*, vol. 6, no. 8, 1997.
- [7] P. L. Silsbee and A. C. Bovik, “Computer lipreading for improved accuracy in automatic speech recognition,” *IEEE Transactions on Speech, and Audio Processing*, vol. 4, no. 5, 1996.
- [8] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, Cambridge, MA: MIT Press, 1997.
- [9] S. Gurbuz, Z. Tufekci, E. K. Patterson, and J. N. Gowdy, “Application of affine-invariant fourier descriptors to lipreading for audio-visual speech recognition,” in *Proceedings of ICASSP*, 2001.
- [10] Z. Tufekci and J.N. Gowdy, “Mel-scaled discrete wavelet coefficients for speech recognition,” in *Proceedings of ICASSP*, 2000.
- [11] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” *IEEE Transactions on Multimedia*, 2000.
- [12] U. Meier, W. Hurst, and P. Duchnowski, “Adaptive bimodal sensor fusion for automatic speechreading,” *Proceedings of ICASSP*, 1996.