

FEATURE SPACE VIDEO STREAM CONSISTENCY ESTIMATION FOR DYNAMIC STREAM WEIGHTING IN AUDIO-VISUAL SPEECH RECOGNITION

Louis H. Terry, Derek J. Shiell, Aggelos K. Katsaggelos

Northwestern University
Department of Electrical Engineering and Computer Science
2145 North Sheridan Road, Evanston, IL, 60208
Email: {lht286,djs470,aggk}@eecs.northwestern.edu

ABSTRACT

Most current audio-visual automatic speech recognition (AV-ASR) systems use static weights to leverage between audio and visual information during information fusion. State of the art research has led to using audio reliability metrics for dynamically changing the fusion weights in order to successfully improve overall recognition results. So far, however, incorporating visual reliability metrics into these audio reliability metric based systems have not significantly improved performance. We introduce a new approach to this problem by inferring the "consistency" between the audio and visual information and leveraging the existing audio reliability metrics to create a video reliability metric. Our approach is formulated in the extracted feature space and, thus, does not rely on analyzing the actual video signal itself. The framework presented in this work competes with the audio-only reliability metric based systems and shows promise to consistently outperform.

Index Terms— Speech Recognition, Hidden Markov Models, Vector Quantization

1. INTRODUCTION

Speech recognition technology continues to spread through modern life to a near ubiquitous role. One finds automatic speech recognition in automated telephone systems, automobiles, computer interfaces, and numerous applications in between. As this expansion unfolds, however, the limits of current speech recognition become apparent; modern systems' performances degrade rapidly in the presence of audio noise [1], and even in clean acoustic environments performance can lag human speech perception by up to an order of magnitude [2]. Only so much signal processing can be done on the incoming audio stream before the returns finally do not outweigh the cost; the need for alternative approaches has become apparent. In recent years visual information has been successfully incorporated into the speech recognition task (for recent reviews, see [3, 4]). While this bimodal approach has seen success, performance still lags behind human perception.

An accurate, robust automatic speech recognition system remains elusive.

An active field of research, audio-visual fusion, explores how to best combine the audio and visual modalities. Currently, audio and visual fusion takes place using empirically pre-determined weights for each modality or using weights dynamically determined by audio reliability metrics such as audio SNR [5]. Many attempts have been made to estimate visual information reliability, among those [6, 7], with little success relative to using only audio reliability metrics. As such, this endeavor remains a fairly open research topic.

We introduce an approach to the visual reliability metric problem utilizing information derived from within the extracted audio and visual feature spaces and the probabilistic relationships between the respective modality's features. Section 2 details the implementation of audio-visual fusion. In Section 3, we present our stream reliability/consistency estimation technique, and in Section 4 we describe how to transform the stream reliability/consistency estimations into stream (modality) weights for the speech recognition system. Section 5 details the experimental setup and results. Lastly, Section 6 concludes and summarizes our work.

2. AUDIO-VISUAL FUSION

Audio-visual automatic speech recognition (AV-ASR) systems generally begin by independently preprocessing each stream and performing feature extraction. Due to differing input rates, the video is then usually interpolated from 30Hz to the audio rate (90Hz). These features may then be fused before recognition, *early integration*, or may be sent directly to the recognizer for *intermediate* or *late integration* schemes.

Intermediate integration leverages the structure of a multi-stream hidden Markov model (MS-HMM) to allow for varying the weights of each stream from the individual time-frame level to the sub-word unit level and on up to the utterance level. With this amount of control, MS-HMMs have become a major tool for audio-visual integration.

In this work, we perform *intermediate integration* on the

time-frame level (i.e. the shortest time level available). Within the MS-HMM framework, the overall likelihood of some observation O_t is modeled as the product of the likelihoods of each stream raised to a power. The stream’s exponent is thus the stream weight. Equation 1 shows this relationship:

$$P(O_t|\lambda) = \prod_{s \in \{1, \dots, N\}} P(O_{s,t}|\lambda)^{\gamma^s}, \quad (1)$$

where O_t represents the complete data observation at time t , $O_{s,t}$ represents the observation of stream s at time t , λ represents the parameters of the recognizer, N is the number of streams, and γ is the stream weight.

3. STREAM RELIABILITY ESTIMATION

Determining the reliability of an audio signal for the purposes of speech recognition has been a well studied problem (see [5] and references therein). However, determining a video stream reliability metric poses many challenges. Among the factors that may influence the reliability of a video sequence for speech recognition are the following:

Visual Tracking: Tracking of the face and/or salient features, such as the lips

Environmental Effects: Lighting changes, occlusions, shadows, etc.

Speaker Appearance: Facial hair, pose, etc.

Video Characteristics: Frame rate, compression artifacts, resolution, etc.

A good video reliability metric based on the video sequence itself must somehow take into account all of these reliability influences. As an alternative, we propose basing a video reliability metric on the *extracted video features* rather than the video sequence.

The proposed system assumes the existence of visually and auditorily ”clean” data. Given this clean data, we extract features as usual, and send them through a vector quantizer with memory to obtain quantization states q_t^A and q_t^V at time t for the audio and video streams, respectively. During training, we then build a conditional probability mass function (PMF) $P(q_t^V|q_t^A)$. In conjunction with an audio stream reliability metric, such as audio SNR, we use the conditional PMF to determine the audio and visual stream weights at time t .

By taking such an approach, we seek to exploit examples of ”consistent” audio-visual features to help identify ”inconsistent” audio-visual parameters. Section 3.1 describes the vector quantizer with memory. Sections 3.2 and 3.3 detail the audio and video stream reliability estimations, respectively. Section 4 will describe the transform between stream reliabilities and stream weights.

3.1. Vector Quantization with Memory

Vector quantization of the audio-visual features not only allows for ease of PMF learning, but has grounding in actual speech production. For instance, if a speaker produces a ’z’ sound, the speaker’s mouth should not be in an open state. Thus, the conditional probability of an open mouth given a ’z’ sound should be very near zero. Because of this correlation between audio production and visual articulation, a quantization approach seems to make sense. Additionally, there is a definite temporal correlation between the possible ”state” of the audio at time t and time $t - 1$ and between the ”state” of the visual articulators at time t and time $t - 1$. For this reason, we choose a vector quantizer with memory (VQwM).

To implement the VQwM we use one independent ergodic HMM (EHMM) per stream. This setup allows for Markovian temporal dependency. For an N state EHMM, each state is labeled from 0 to $N - 1$ and the state occupied at time t becomes the quantized label (q_t^A or q_t^V , for audio and video, respectively).

Training the VQwM begins by using k-means to cluster the individual observations. These clusters are modeled using Gaussian Mixture Models and become the states of the EHMM. An initial estimate of the HMM parameters are constructed by directly assigning each observation in each sequence to the state into which it was clustered. A second pass through the training data refines the initial estimates, where each observation sequence is used as input to the EHMM and the state transitions recorded and updated.

3.2. Audio Stream Reliability Metric

In most high SNR conditions, more information is contained in the audio stream than the video stream. That is not to say that the video plays no importance, but that the complimentary information contained in the video stream usually is more helpful than the video information absent the audio. Thus, the more reliable the audio, the more we should value the audio relative to the video. Given the availability of informative audio reliability metrics, we utilize the SNR as our metric of choice.

3.3. Video Stream Consistency Estimation

During the training phase, we accumulate $P(q_t^V|q_t^A)$ after the audio and visual features have been quantized. During recognition, we use this PMF as a look-up table to determine the quantitative ”consistency” of the video given the audio. We now make the argument that four cases of qualitative consistency exist:

Consistent AV, Reliable Audio: Given that the audio is reliable and that the likelihood of the observed visual state is high given the observed audio, the video is reliable.

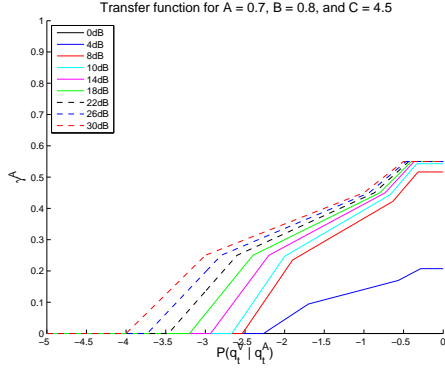


Fig. 1. Transfer function for $A = 0.7$, $B = 0.8$, and $C = 4.5$

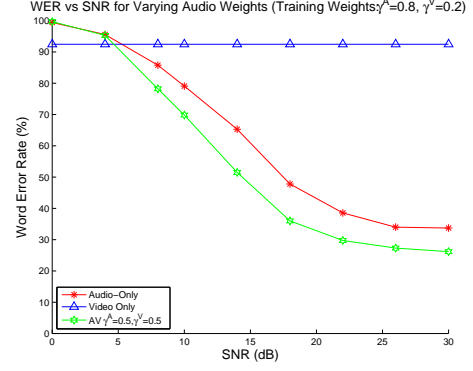


Fig. 2. Baseline AV static stream weights, audio-only, and video-only results

Inconsistent AV, Reliable Audio: Given that the audio is reliable and that the likelihood of the observed visual state is low given the observed audio, the video is unreliable.

Consistent AV, Unreliable Audio: Given that the audio is unreliable and that the likelihood of the observed visual state is high given the observed audio, we can not accurately deduce the reliability of the video.

Inconsistent AV, Unreliable Audio: Given that the audio is unreliable and that the likelihood of the observed visual state is low given the observed audio, we can not accurately deduce the reliability of the video.

By interpreting these four cases, we can transform the visual consistency and the audio reliability to stream weights.

4. FROM STREAM RELIABILITIES TO STREAM WEIGHTS

Given SNR as our audio stream reliability metric and the video consistency metric, we define the transform

$$\mathcal{T} : \{P(q_t^V | q_t^A), SNR_t\} \rightarrow \{\gamma_t^A, \gamma_t^V\}, \text{ s.t. } \gamma_t^A + \gamma_t^V = 1. \quad (2)$$

We specify γ_t^A monotonically non-decreasing as SNR_t increases. For dynamic stream weighting based on audio metrics, a sigmoid is a common function to use when mapping $SNR_t \rightarrow \gamma_t^A$ [8], though statistical methods have been proposed [6]. For simplicity, we use a sigmoid function multiplied by an exponential in the SNR_t dimension. In the $P(q_t^V | q_t^A)$ dimension we define a piecewise linear function at 0dB and at 30dB to be interpolated to intermediate values using the sigmoid-exponential function in SNR_t . Equation 3 details the form of the sigmoid-exponential function,

$$1 - \left[(1 - e^{-A * SNR_t}) * \frac{1}{1 + e^{-B * (SNR_t - C)}} \right], \quad (3)$$

where A controls the rate of the exponential, B controls the shape of the sigmoid, and C controls the offset of the center of the sigmoid. Figure 1 displays the final transfer function for each SNR.

5. EXPERIMENTAL RESULTS

As a preliminary test of our proposed method, we use the Bernstein Database [9] to test our method's success in a large vocabulary continuous speech recognition (LVCSR) context. LVCSR experiments constitute the most difficult test of a recognition system and emulate real-world situations for the use of ASR systems.

For a baseline, we use the system presented in [10] which also uses the Bernstein Database. Mel-frequency cepstral coefficients (MFCC) and MPEG-4 compliant Facial Animation Parameters (FAPs) are used for audio and visual features, respectively. A three-emitting state left-to-right triphone based HMM is used for recognition. Visual features are upsampled to the audio rate of 90Hz, allowing the dynamic stream weights to change every $1/90^{th}$ of a second.

As in [10], the first 379 TIMIT sentences in the database are used for training. In this work they are used for both training the conditional PMF as well as training the speech recognizer. Firstly, the conditional PMF is determined, and, subsequently, the PMF is used for dynamic stream weighting on those same sentences during training the HMM. The remaining 85 TIMIT sentences are used for recognition using dynamic stream weights.

White noise was added to the training sentences to produce the "clean" audio at an SNR of 30dB. For the testing data, the noise was added to achieve SNRs of 0, 4, 8, 10, 14, 18, 22, 26, and 30db. Each experiment consisted of training the PMF and recognizer, then testing that recognizer on data from each of the SNRs. The exact SNR was given to the stream weight computation system. The experiment is run ten

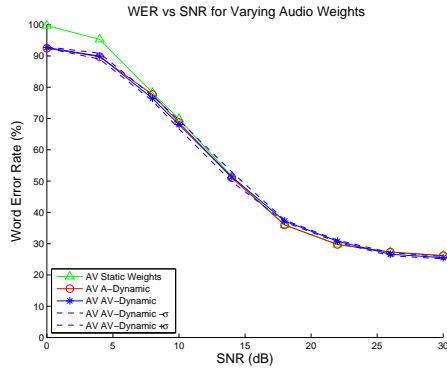


Fig. 3. Baseline AV static stream weights, AV audio-only reliability dynamic stream weights, and AV AV-reliability dynamic stream weights

times due to variations in the VQwM introduced by the slight randomness of the clusters output by the k-means algorithm (this randomness is owed to the random initialization of the algorithm). These results are then averaged.

Figure 2 shows the baseline (static) audio-visual, the audio-only, and video-only results. Figure 3 compares our proposed approach (+/- one standard deviation) to the static audio-visual case ($\gamma^A = \gamma^V = 0.5$) and the audio-visual case using audio-only reliability metrics to drive the dynamic stream weights (Figure 4 is an enlarged view of figure 3).

As shown in these figures, on average, our proposed method performs competitively with the audio-only reliability system. However, within one standard deviation, our system shows the ability to outperform the audio-only system.

6. SUMMARY

This work presents a new framework and approach to the visual stream reliability metric problem for use with dynamic stream weighting in automatic speech recognition. We detail the proposed approach and present initial results on par with other state of the art systems that incorporate visual stream reliability.

7. REFERENCES

- [1] Y. Gong, “Speech recognition in noisy environments: a survey,” *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.
- [2] R. Lippmann, “Speech perception by humans and machines,” *Speech Communication*, vol. 22, pp. 1–15, 1997.
- [3] P. S. Aleksic, G. Potamianos, and A. K. Katsaggelos, *Handbook of Image and Video Processing*, chapter Ex-

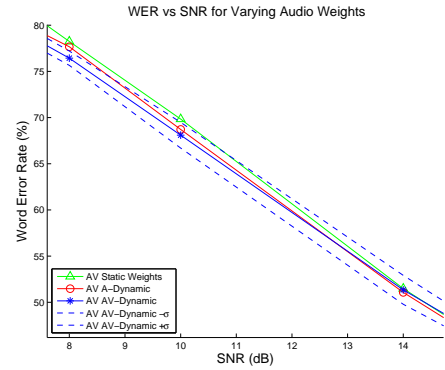


Fig. 4. Baseline AV static stream weights, AV audio-only reliability dynamic stream weights, and AV AV-reliability dynamic stream weights (enlarged view of figure 3)

ploiting visual information in automatic speech processing, pp. 1263–1289, Academic Press, 2005.

- [4] G. Potamianos, C. Neti, J. Luetttin, and I. Matthews, “Audio-visual automatic speech recognition: An overview,” in *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds. 2004, MIT Press.
- [5] H. Glotin, D. Vergyr, C. Neti, G. Potamianos, and J. Luetttin, “Weighting schemes for audio-visual fusion in speech recognition,” *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP ’01). 2001 IEEE International Conference on*, vol. 1, pp. 173–176 vol.1, 2001.
- [6] E. Marcheret, V. Libal, and G. Potamianos, “Dynamic stream weight modeling for audio-visual speech recognition,” *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV-945–IV-948, 15-20 April 2007.
- [7] V. Pitsikalis, A. Katsamanis, G. Papandreou, and P. Maragos, “Adaptive multimodal fusion by uncertainty compensation,” *INTERSPEECH 2006*, 2006.
- [8] M. Heckmann, F. Berthommier, and K. Kroschel, “Noise adaptive stream weighting in audio-visual speech recognition,” *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1260–1273, 2002.
- [9] L. Bernstein and S. Eberhardt, “Johns hopkins lipreading corpus i-ii,” Tech. Rep., Johns Hopkins University, Baltimore, Md, USA, 1986.
- [10] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, “Audio-visual speech recognition using mpeg-4 compliant visual features,” *EURASIP Journal on Applied Signal Processing*, pp. 1213–1227, 2002.