

# STREAM CONFIDENCE ESTIMATION FOR AUDIO-VISUAL SPEECH RECOGNITION

Gerasimos Potamianos

Chalapathy Neti

IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.

email: {gpotam, cneti}@us.ibm.com

## ABSTRACT

We investigate the use of single modality confidence measures as a means of estimating adaptive, local weights for improved audio-visual automatic speech recognition. We limit our work to the toy problem of audio-visual phonetic classification by means of a two-stream Gaussian mixture model (GMM), where each stream models the class conditional audio- or visual-only observation probability, raised to an appropriate exponent. We consider such stream exponents as two-dimensional piecewise constant functions of the audio and visual stream local confidences, and we estimate them by minimizing the misclassification error on a held-out data set. Three stream confidence measures are investigated, namely the stream entropy, the n-best likelihood ratio average, and an n-best stream likelihood dispersion measure. The later results in superior audio-visual phonetic classification, as indicated by our experiments on a 260-subject, 40-hour long, large vocabulary, continuous speech audio-visual dataset. By using local, dispersion-based stream exponents, we achieve an additional 20% phone classification accuracy improvement over the improvement that global stream exponents add to clean audio-only phonetic classification. The performance of the algorithm however still falls significantly short of an “oracle” (cheating) confidence estimation scheme.

## 1. INTRODUCTION

It is well known that humans fuse information from both the audio and visual stimuli to recognize speech [1], as well as that the visual modality contains some complementary speech information to the audio one [2]. Not surprisingly, *automatic speech recognition* (ASR) by using the video sequence of the speaker’s lips, namely *automatic lipreading*, or *speechreading*, has been shown to improve traditional audio-only ASR performance over a wide range of conditions [3]-[10]. Of course, such an improvement greatly depends on the audio-visual fusion strategy used.

Various fusion techniques have been recently proposed in the literature for audio-visual ASR [10]. Among those, the most commonly used method assumes that the class conditional score of the audio-visual feature vector is a weighted average of the log-likelihoods of each single modality observation vector (audio- and visual-only) [4-9]. This approach corresponds to a popular classifier fusion technique (adaptive weighting/product rule [11]), and in ASR, it gives rise to the *multi-stream hidden Markov model* (HMM) representation, which has recently been used, among others, for multi-band audio-only ASR [12], [13].

To the best of our knowledge, multi-stream HMM based audio-visual fusion methods have limited the stream exponents to constant values (possibly class dependent) over an entire dataset [7]-[9], or over an entire utterance [4]-[6]. Our intent is to adaptively estimate such exponents at a finer temporal level, allowing them to vary within the utterance of interest. In this paper, we investigate local stream exponent estimation in the toy

problem of audio-visual phonetic classification, by means of the multi-stream *Gaussian mixture model* (GMM) classifier. Single-stream (audio- and visual-only) GMMs are initially trained for each phone (class), based on an available training set. The stream exponents are then assumed to be piecewise constant functions of the two-dimensional audio-visual “confidence” vector, and are estimated by minimizing the misclassification error on a separate held-out data set.

The following are of interest in the above approach: (a): The choice of appropriate stream (modality) confidence measures; (b): The partitioning of the two-dimensional confidence space into “confidence bins”; and (c): The estimation of stream exponents, given held-out data within their confidence partition. In this paper, we mainly address the first issue. In particular, we investigate three stream confidence measures, namely the stream entropy [4], [13], the n-best likelihood ratio average, and an n-best stream likelihood dispersion measure [5], [6]. We subsequently estimate GMM stream exponents based on these measures. We conduct experiments to evaluate the performance of the resulting multi-stream classifier using a 260-subject, 40-hour long, large vocabulary, continuous speech audio-visual database.

The paper is structured as follows: Section 2 introduces necessary notation and reviews the multi-stream GMM, Section 3 defines a number of confidence measures, and Section 4 describes stream exponent estimation, based on stream confidences. Section 5 is devoted to a brief overview of the speechreading system, while Section 6 presents our audio-visual phonetic classification experiments. Finally, Section 7 summarizes our findings.

## 2. THE MULTI-STREAM GMM CLASSIFIER

Let us denote the *time-synchronous* audio-visual feature observation sequence that corresponds to a spoken utterance by<sup>1</sup>

$$\{ \underline{Q}^{(t)} = [Q_A^{(t)}, Q_V^{(t)}] \in \mathbb{R}^D, 1 \leq t \leq T \}, \quad (1)$$

where  $Q_s^{(t)} \in \mathbb{R}^{D_s}$ , for  $s = A, V$ , represent the single modality (audio- and visual-only) feature vectors, and  $D = D_A + D_V$  is the bimodal feature vector size. Based on  $\underline{Q}^{(t)}$ , we are interested in classifying each time instant  $t$  of the utterance as belonging to one of  $|\mathcal{C}|$  possible phonetic classes  $c \in \mathcal{C} = \{1, \dots, |\mathcal{C}|\}$ . In this work, 52 such classes are considered, as in [14].

Each  $Q_s^{(t)}$ ,  $s = A, V$ , contains relevant information about  $c^{(t)}$ , the phonetic class at time  $t$ . We capture this by assuming a *Gaussian mixture model* (GMM) as the single modality class conditional observation probability, namely

$$Pr(Q_s^{(t)} | c) = \sum_{m=1}^{M_{cs}} w_{c m s} \mathcal{N}_{D_s}(Q_s^{(t)}; \underline{\mu}_{c m s}, \underline{\sigma}_{c m s}), \quad (2)$$

<sup>1</sup>With some abuse of notation,  $T$  will also denote the total number of feature vectors (observations) in a set of utterances.

for all  $c \in \mathcal{C}$ , and  $s = A, V$ . In (2), *mixture weights*  $w_{cms}$  are positive adding up to one,  $M_{cs}$  denotes the number of class  $c$  mixtures for stream  $s$ , and  $\mathcal{N}_D(\underline{y}; \underline{\mu}, \underline{\sigma})$  is the  $D$ -variate normal distribution with mean  $\underline{\mu}$  and diagonal covariance  $\underline{\sigma}$ . GMM parameters

$$\underline{\theta}_s = [ (w_{cms}, \underline{\mu}_{cms}, \underline{\sigma}_{cms}), m = 1, \dots, M_{cs}, c \in \mathcal{C} ],$$

for  $s = A, V$ , can be computed by means of the *expectation-maximization* (EM) algorithm [15].

The *maximum-a-posteriori* (MAP) class estimate of  $c^{(t)}$ , based on the single modality observation  $\underline{Q}_s^{(t)}$ , can then be obtained as

$$\hat{c}_s^{(t)} = \arg \max_{c \in \mathcal{C}} \{ Pr(\underline{Q}_s^{(t)} | c) Pr(c) \}, \quad (3)$$

where, for simplicity a uniform prior  $Pr(c) = 1/|\mathcal{C}|$  is assumed. Finally, the frame level phonetic *misclassification error* is computed by comparing at every instance of  $t$ , the correct phone label  $c^{(t)}$  to its MAP estimate (3), and it equals

$$MCE_s = \frac{1}{T} \sum_{t=1}^T (1 - \delta(c^{(t)}, \hat{c}_s^{(t)})), \quad (4)$$

where  $\delta(x, y) = 1$  (0), iff  $x = y$  ( $x \neq y$ ).

We now model the class conditional bimodal observation probability by considering a two-stream GMM, namely by assuming that (see also (1), (2))

$$Sc[\underline{Q}^{(t)} | c] = \prod_{s \in \{A, V\}} \left[ \sum_{m=1}^{M_{cs}} w_{cms} \mathcal{N}_{D_s}(\underline{Q}_s^{(t)}; \underline{\mu}_{cms}, \underline{\sigma}_{cms}) \right]^{\gamma_s^{(t)}}, \quad (5)$$

where  $\gamma_s^{(t)}$  are time-dependent stream exponents, that locally model the reliability of each modality (stream). In this work, they are constrained to satisfy

$$0 \leq \gamma_A^{(t)}, \gamma_V^{(t)} \leq 1, \quad \gamma_A^{(t)} + \gamma_V^{(t)} = 1, \quad \text{for all } t = 1, \dots, T. \quad (6)$$

Note that (5) does not represent a probability mass function in general, therefore we refer to it as a *score*. Similarly to (3), the MAP estimate of  $c^{(t)}$ , based on the bimodal observation  $\underline{Q}^{(t)}$  is

$$\hat{c}^{(t)} = \arg \max_{c \in \mathcal{C}} \{ Sc[\underline{Q}^{(t)} | c] \}, \quad (7)$$

under the uniform class prior assumption. In addition, and similarly to (4), the bimodal misclassification error is

$$MCE = \frac{1}{T} \sum_{t=1}^T (1 - \delta(c^{(t)}, \hat{c}^{(t)})). \quad (8)$$

One hopes that the multi-stream classification decisions (7) are on the average better than both single modality ones (3), and thus, that  $MCE \leq \min\{MCE_A, MCE_V\}$  holds.

The focus of this paper is the choice of exponents  $\gamma_s^{(t)}$  such that MCE becomes “small”. It is reasonable to expect that “good” choices of such exponents should be functions of the confidence of the single stream classifiers (2) on the class decision, based on the single modality observation, which we shall denote by  $\mathcal{I}_s^{(t)} = g(\underline{Q}_s^{(t)}, \underline{\theta}_s)$ . We namely assume that  $(\gamma_A^{(t)}, \gamma_V^{(t)}) = f(\mathcal{I}_A^{(t)}, \mathcal{I}_V^{(t)})$ , subject to (6). Two issues are clearly of interest in this approach: The choice of confidence measures  $\mathcal{I}_s^{(t)}$ , and the design of function  $f$ . These issues are addressed in the following two sections.

### 3. STREAM CONFIDENCE MEASURES

Let us denote by  $c_{s,n}^{(t)}$ ,  $n = 1, \dots, N$ , the ranked  $N$ -best phone class decisions of single stream classifier (2), given the single stream observation  $\underline{Q}_s^{(t)}$  (note that  $\hat{c}_s^{(t)} = c_{s,1}^{(t)}$ ). Let us also denote the log-likelihoods of the  $n$ th-best and of the correct hypothesis by  $\mathcal{R}_{s,n}^{(t)} = \log Pr(\underline{Q}_s^{(t)} | c_{s,n}^{(t)})$ , and  $\mathcal{F}_s^{(t)} = \log Pr(\underline{Q}_s^{(t)} | c^{(t)})$ , respectively. Notice that  $\mathcal{F}_s^{(t)} \leq \mathcal{R}_{s,1}^{(t)}$ , and that  $\mathcal{R}_{s,n}^{(t)} \geq \mathcal{R}_{s,n+1}^{(t)}$ , for  $n = 1, \dots, |\mathcal{C}| - 1$ . The following represent some natural choices for single stream confidence measures:

**Stream entropy.** We consider the negative entropy of the class posterior probability mass function, given by

$$\mathcal{I}_{s,E}^{(t)} = - \frac{\sum_{n=1}^{|\mathcal{C}|} \mathcal{R}_{s,n}^{(t)} \exp[\mathcal{R}_{s,n}^{(t)}]}{\sum_{n=1}^{|\mathcal{C}|} \exp[\mathcal{R}_{s,n}^{(t)}]} + \log \sum_{n=1}^{|\mathcal{C}|} \exp[\mathcal{R}_{s,n}^{(t)}]. \quad (9)$$

Clearly, values of  $\mathcal{I}_{s,E}^{(t)} \approx 0$  indicate strong confidence on the specific stream observation by the GMM classifier, whereas values  $\mathcal{I}_{s,E}^{(t)} \approx \log |\mathcal{C}|$  indicate lack of any discrimination among the various classes by the classifier. Notice that (9) has been proposed in [13] in the context of multi-band audio-only ASR, as well as in [4] for audio-visual speech recognition.

**Average N-best log-likelihood difference.** The likelihood ratios between the first  $N$  classification decisions are informative about the class discrimination based on the observation. A reasonable choice for capturing such discrimination is to use the log-likelihood difference average between the  $N$ -best classification decisions. Such confidence measure is given by

$$\mathcal{I}_{s,L}^{(t)} = \frac{1}{N-1} \sum_{n=2}^N (\mathcal{R}_{s,1}^{(t)} - \mathcal{R}_{s,n}^{(t)}), \quad (10)$$

where  $N \geq 2$ . Clearly, “large” values of  $\mathcal{I}_{s,L}^{(t)}$  indicate high classification decision confidence.

**N-best log-likelihood dispersion.** An alternative to the above measure, suggested in [5], is the log-likelihood “dispersion” among the  $N$ -best classifier decisions. This is given by

$$\mathcal{I}_{s,D}^{(t)} = \frac{2}{N(N-1)} \sum_{n=1}^N \sum_{n'=n+1}^N (\mathcal{R}_{s,n}^{(t)} - \mathcal{R}_{s,n'}^{(t)}), \quad (11)$$

where  $N \geq 2$ . As with (10), “large” values of  $\mathcal{I}_{s,D}^{(t)}$  indicate high confidence in the decision of the classifier in question.

**Misclassification discriminant function.** Finally, it is of interest to consider the smooth function of the misclassification error used in discriminative training algorithms such as the *generalized probabilistic descent* (GPD) [15]. This is given by

$$\mathcal{I}_{s,M}^{(t)} = \left[ 1 + \frac{\exp[\mathcal{F}_s^{(t)}]}{\frac{1}{N_t} \sum_{n=1}^N \delta(c_{s,n}^{(t)}, c^{(t)}) \exp[\mathcal{R}_{s,n}^{(t)}]} \right]^{-1}, \quad (12)$$

where  $N_t = \sum_{n=1}^N \delta(c_{s,n}^{(t)}, c^{(t)})$ . Note that (12) is a “cheating” measure on the test set, as it requires knowledge of the correct class  $c^{(t)}$ .

### 4. STREAM EXPONENT ESTIMATION

Given a choice for the stream confidence measures  $\mathcal{I}_s^{(t)}$ ,  $s = A, V$ , and an available held-out data set, the question becomes how to estimate an optimal function  $(\gamma_A^{(t)}, \gamma_V^{(t)}) = f(\mathcal{I}_A^{(t)}, \mathcal{I}_V^{(t)})$ ,

that minimizes the misclassification error (8) on the held-out set, subject to (6), and, in addition, it has the ability to generalize to unseen data. In this paper, we adopt a simple, suboptimal approach to stream exponent estimation. Our method is based on partitioning the two-dimensional “confidence” space (a subset of  $\mathbb{R}^2$ ) in bins, and subsequently estimating a *piecewise constant* function  $f$  on these bins.

The algorithm first constructs two single stream confidence partitions by considering equally spaced bins in the intervals  $[\max\{\min_s, \mu_s - 2.5\sigma_s\}, \mu_s]$  and  $[\mu_s, \min\{\max_s, \mu_s + 2.5\sigma_s\}]$ , where  $\min_s = \min_t \mathcal{I}_s^{(t)}$ ,  $\max_s = \max_t \mathcal{I}_s^{(t)}$ , and  $\mu_s, \sigma_s$ , denote the stream confidence measure sample mean and variance, all computed over the held-out data set. Confidences outside these intervals are mapped to their closest partition. The two-dimensional partition is set to the “product” of the two single stream partitions. The method thus results to a partitioning  $\mathcal{B} = \{b_1, b_2, \dots, b_{|\mathcal{B}|}\}$  of  $\mathbb{R}^2$ . Of course, alternative partitioning methods can be considered, for example, by means of histogram equalization of the confidence measure values on the held-out set.

Given the derived partition, each time instant  $t$  is mapped to a confidence bin  $b(t) \in \mathcal{B}$ , such that  $(\mathcal{I}_A^{(t)}, \mathcal{I}_V^{(t)}) \in b(t)$ . Constant values of  $(\gamma_A, \gamma_V)$  over that bin are computed by minimizing  $\sum_{t: b(t) \in b_j} (1 - \delta(\hat{c}^{(t)}, \hat{c}^{(j)}))$  (see also (8)), subject to (6). Such minimization is explicitly carried out by searching a fine grid of points that satisfy (6). Clearly, this amounts to a one-dimensional search for one of the two exponents in the interval  $[0, 1]$ . Here, we use 101 equally spaced points in this interval. Alternative optimization methods, such as the GPD algorithm [15] can be used instead.

## 5. THE SPEECHREADING SYSTEM

Before reporting our database experiments, a brief description of our automatic speechreading system is warranted [14]. Audio and visual feature vectors are extracted from the audio-waveform within a sliding window, or from the video pixel values within an extracted *region of interest* (ROI), respectively. In both cases, a similar three-stage cascade algorithm is used: First, a signal processing data transform is applied, followed by a *linear discriminant analysis* (LDA) projection on a number of consecutive transformed vectors. Finally, a *maximum likelihood linear transform* (MLLT) results in a final data rotation. The extracted audio features are 60-dimensional, whereas the extracted visual ones are of dimension 41.

In more detail, in the case of visual feature extraction, a face tracking algorithm is first employed to find the location and size of the subject’s mouth region [14]. A  $64 \times 64$  pixel, size normalized ROI is then extracted at every video frame, at the rate of 60 Hz. A separable, two-dimensional discrete cosine transform (DCT) is subsequently applied to the ROI, and the 24 highest energy (over the entire database) transform coefficients are retained. This 24-dimensional feature vector is interpolated to 100 Hz (the audio feature rate), so as to provide time synchronous features to the audio ones. Feature mean normalization over the entire sequence is also applied element-wise to the interpolated feature vector, in order to eliminate variations due to lighting, among others. Fifteen consecutive such feature vectors are then concatenated and projected to a lower dimensional space (of size 41) by means of LDA. Finally, the MLLT data rotation is applied [14].

## 6. DATABASE AND EXPERIMENTS

We have been collecting a multi-subject, continuous, large vocabulary, audio-visual database, using IBM ViaVoice training utterance scripts. Currently, it consists of 260 subjects and close to 40 hours of speech. The database contains full frontal face color video of the subjects with minor face-camera distance and

Method	Acc.	Method	Acc.
Audio-only	50.38	AV-confidence (9)	54.44
Visual-only	28.34	AV-confidence (10)	55.05
AV-baseline	54.35	AV-confidence (11)	55.19
AV-confidence (12)	59.88	AV-subject depend.	54.43

**Table 1:** Audio-, visual-only and multi-stream audio-visual test set phonetic classification accuracy (%). The baseline system corresponds to two globally set exponents. Subject dependent exponent performance is also depicted.

lighting variations. The video is captured at a resolution of  $704 \times 480$  pixels (interleaved), a frame rate of 60 Hz, and is MPEG2 encoded, whereas the audio is captured at 16 KHz. Three sets are constructed from this database: A 35-hour training set (about 17,000 utterances), a 4-hour held-out set (about 2,000 utterances), and a 1-hour test set (close to 400 utterances). Data from all 260 subjects are present in all three sets, therefore this data partitioning corresponds to a *multi-subject* training/testing scenario.

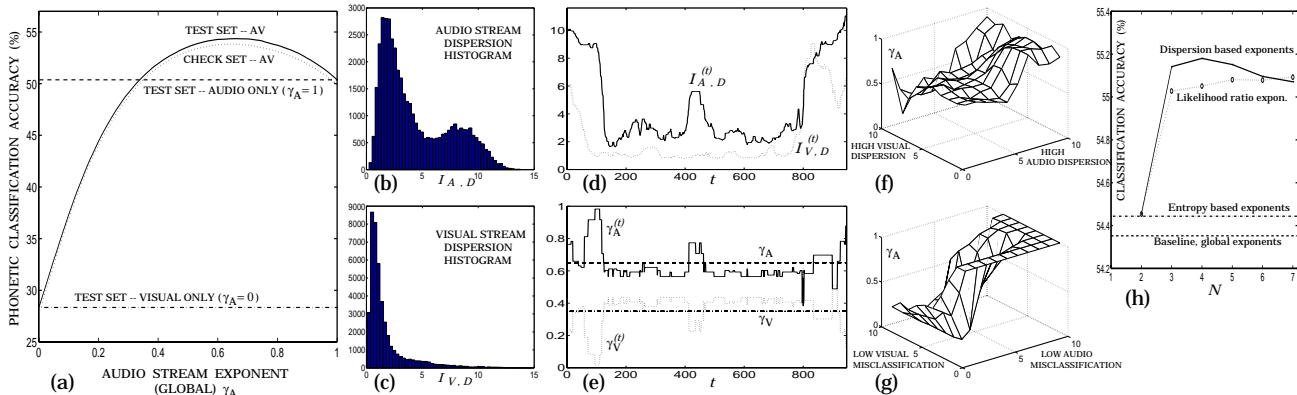
For fast experimentation, we train single stream GMM classifiers with only 5 Gaussian mixtures for each one of the 52 phonetic classes of interest [14]. We use the EM algorithm for this task, and a fixed segmentation of the training utterances into classes obtained by forced alignment using an audio-only HMM [14]. The audio-only and visual-only GMM phonetic classification accuracies on the test set are depicted in Table 1, and are 50.38% and 28.34% respectively. Notice that significantly higher phonetic classification accuracies can be achieved by using more mixtures, prior class information, or phone segment duration information (see [14], where a 80.52% audio-only and a 48.85% visual-only accuracy are reported for a similar task).

We subsequently proceed to optimize stream exponents on the held-out data set. Assuming one global exponent per stream for the entire database, we obtain an audio-visual phonetic classification accuracy of 54.35%. We characterize this as the “baseline” system, and we consider the various stream confidence measures of Section 3, aiming at improving classification performance. For each confidence measure, 100 bins are considered, where stream exponents are estimated. In addition, while testing, confidences are smoothed by means of *median* filtering over a interval of 25 consecutive frames. As depicted in Table 2, the best performance is obtained by the  $N$ -best log-likelihood dispersion confidence which results in 55.19% audio-visual phonetic classification accuracy, when using  $N = 4$ . Notice that this corresponds to an additional 21% relative improvement w.r.t. the improvement that the baseline achieves over the audio-only performance.

The  $N$ -best likelihood ratio (10) performs slightly worse than the dispersion (11) (results in a 55.05% accuracy), whereas the entropy (9) fails to significantly improve performance over the baseline system, giving only a 54.44% accuracy. It is interesting to note that the “cheating” misclassification measure (12) results in a 59.88% accuracy, indicating that there still exists significant room for improvement.

Since our testing scenario is multi-subject, we can estimate optimal weights per subject collected. Such exponents are optimized on the part of the held-out data set that corresponds to the specific subject. Surprisingly, this method gives a very small improvement over the baseline system, namely an accuracy of 54.43%.

It is worth noticing that temporal smoothing of the stream confidences helps performance, when median filtering is used. For example, in the case of the dispersion based stream exponent estimation, the performance drops to 54.94%, when no temporal



**Figure 1:** (a): Test and held-out set audio-visual phonetic classification accuracy using global stream exponents. (b,c): Audio and visual-only  $N$ -best dispersion histogram on the held-out set. (d,e): Dispersion and estimated stream exponents for a database utterance. (f,g): Estimated audio exponents per bin of dispersion (11) or misclassification (12) based confidence. (h): Audio-visual phonetic classification accuracy on the test set for various confidence measures.

smoothing is employed. Furthermore, using utterance-wide constant stream exponents based on the mean of the two dispersions over the entire sequence hurts performance significantly, by further reducing accuracy to 54.37%.

In Fig. 1, a number of facts concerning confidence based stream exponent estimation are depicted. Fig. 1(a) shows audio-visual accuracy on the test and held-out sets versus the choice of a global audio-stream exponent. Notice that both curves are quite flat close to their peaks, which shows that approximate estimation of exponents suffices. In addition, the peaks of the two curves coincide, which encourages estimation of such parameters on a held-out data set that matches the test set. Figs. 1(b,c) depict the dispersion histogram on a small subset of the held-out data set (50 sequences). Fig. 1(d) shows smoothed dispersion measures over a typical database utterance. The peak areas of the audio dispersion are during silence intervals. Fig. 1(e) shows that within such intervals the audio modality is in general preferred. However, in speech intervals, the visual stream exponent is somewhat increased over its global value over the entire dataset. Figs. 1(f,g) plot the estimated audio stream w.r.t. the two-dimensional bins in the dispersion or misclassification confidence space. Not surprisingly, the second measure obtains more "extreme" estimates of the exponents (further away from their global database values). Finally, Fig. 1(h) depicts classification accuracy of the dispersion and likelihood ratio based methods, w.r.t. the value of  $N$  in (10), (11). Notice that the dispersion metric is optimal around  $N = 4$ . The baseline and entropy-based system performances are also shown.

## 7. SUMMARY

We investigated the use of single modality confidence measures as a means of estimating adaptive, local weights for improved audio-visual automatic speech recognition. We have limited our work to the toy problem of audio-visual phonetic classification by means of a two-stream Gaussian mixture model (GMM), and we have obtained significant classification improvement by using a log-likelihood dispersion based stream confidence estimation scheme, over a baseline system of constant stream exponents. In view of these encouraging results, we are currently exploring means of expanding this work into large vocabulary, continuous ASR in the setting of the Johns Hopkins Summer Workshop 2000, in Baltimore, Maryland. Such work is in progress and will be reported in the near future.

## 8. REFERENCES

1. H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, 1976.
2. D.W. Massaro and D.G. Stork, "Speech recognition and sensory integration," *American Scientist*, vol. 86, pp. 236-244, 1998.
3. C. Bregler and Y. Konig, "'Eigenlips' for robust speech recognition," *Proc. Int. Conf. Acoust. Speech Signal Process.*, Adelaide, pp. 669-672, 1994.
4. I. Matthews, J.A. Bangham, and S. Cox, "Audio-visual speech recognition using multiscale nonlinear image decomposition," *Proc. Int. Conf. Speech Lang. Process.*, Philadelphia, pp. 38-41, 1996.
5. A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 461-471, 1996.
6. A. Rogozan, P. Deléglise, and M. Alissali, "Adaptive determination of audio and visual weights for automatic speech recognition," *Proc. Europ. Tut. Works. Audio-Visual Speech Process.*, Rhodes, pp. 61-64, 1997.
7. P. Jorlin, "Word dependent acoustic-labial weights in HMM-based speech recognition," *Proc. Europ. Tut. Works. Audio-Visual Speech Process.*, Rhodes, pp. 69-72, 1997.
8. J. Luetttin, "Towards speaker independent continuous speechreading," *Proc. Eurospeech*, Rhodes, pp. 1991-1994, 1997.
9. G. Potamianos and H.P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," *Proc. Int. Conf. Acoust. Speech Signal Process.*, Seattle, pp. 3733-3736, 1998.
10. M.E. Hennecke, D.G. Stork, and K.V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 331-349, 1996.
11. A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 4-37, 2000.
12. H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," *Proc. Int. Conf. Speech Lang. Process.*, Philadelphia, pp. 426-429, 1996.
13. S. Okawa, T. Nakajima, and K. Shirai, "A recombination strategy for multi-band speech recognition based on mutual information criterion," *Proc. Europ. Conf. Speech Comm. Tech.*, Budapest, pp. 603-606, 1999.
14. G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, "A cascade image transform for speaker independent automatic speechreading," to appear: *Proc. Int. Conf. Multimedia Expo.*, New York, 2000.
15. L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, 1993.