

Joint Audio-Visual Speech Processing for Recognition and Enhancement

Gerasimos Potamianos, Chalapathy Neti, and Sabine Deligne

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA
{gpotam,cneti,deligne}@us.ibm.com

Abstract

Visual speech information present in the speaker's mouth region has long been viewed as a source for improving the robustness and naturalness of human-computer-interfaces (HCI). Such information can be particularly crucial in realistic HCI environments, where the acoustic channel is corrupted, and as a result, the performance of traditional automatic speech recognition (ASR) systems falls below usability levels. In this paper, we review two general approaches that utilize visual speech to improve ASR in acoustically challenging environments: One directly combines features extracted from the acoustic and visual channels, aiming at superior recognition performance of the resulting audio-visual ASR system. The other seeks to eliminate the noise present in the acoustic features, aiming at their audio-visual based enhancement, and thus resulting in improved speech recognition. We present a number of techniques recently introduced in the literature for bimodal ASR and enhancement, and we study their performance using a suitable audio-visual database. Among the methods considered, our recognition experiments demonstrate that decision based combination of audio and visual features significantly outperforms simpler feature based integration methods for audio-visual ASR. For audio feature enhancement, a non-linear technique is more successful than a regression-based approach. As expected, bimodal ASR and enhancement outperform their audio-only counterparts.

1. Introduction

Human speech is by nature bimodal, both in its production and perception [1, 2]. For example, the visual modality benefit to speech intelligibility in noise has been quantified as far back as in 1954 [3], whereas the fact that humans integrate audio and visual stimuli to perceive speech has been demonstrated by the McGurk effect [4]. The visual channel plays a major role in human-to-human speech communication as it helps speaker localization, contains speech segmental information that supplements the audio, and provides complimentary information about the place of articulation [5, 6]. In addition, a number of studies in the literature have quantified the fact that there exists significant correlation between lower face movements and the produced acoustic signal [7, 8].

Motivated by the above facts, over the past twenty years, researchers have been investigating the integration of the visual modality into the speech channel of the human-computer-interface (HCI), aiming in improving its robustness and naturalness. Indeed, various important HCI components, such as speaker identification, verification [9, 10], and localization [11], speech event detection [12], speech signal separation [13], coding [14], video indexing and retrieval [15], and text-to-speech [16, 17], have been shown to benefit from the visual channel.

The bulk of bimodal speech research however, starting with Petajan's work in [18], has concentrated in the field of *auto-*

matic speech recognition (ASR). ASR represents a crucial HCI component that, in its traditional audio-only form, significantly lags in performance with respect to human speech perception [19]. In addition, it lacks robustness to acoustic degradation, in spite of a number of techniques introduced in the literature to compensate for noise [20, 21, 22]. Visual speech, on the other hand, provides a source of information orthogonal to the audio input, obviously not affected by acoustic noise. Not surprisingly, *audio-visual ASR* (AV-ASR) systems that utilize speech information from both channels have been demonstrated to significantly improve ASR robustness to noise, first for small-vocabulary tasks [18, 23, 24, 25], and, more recently, for *large-vocabulary, continuous speech recognition* (LVCSR) [26, 27].

In an indirect approach to improve ASR performance in noise, the visual modality has also been investigated as a means of noisy acoustic signal or feature *enhancement*. A number of traditional audio-only based enhancement techniques [22, 28] have been recently extended to incorporate visual speech information, starting with the work of Girin, et. al., in [29]. Both linear [30, 31] and non-linear [30, 32] methods have been shown to successfully enhance audio features corrupted by noise.

In this paper, we review audio-visual based ASR and audio feature enhancement, and we compare the speech recognition performance of a number of representative techniques on a suitable audio-visual database. In more detail, Section 2 gives a brief introduction on speech informative visual feature extraction, with particular emphasis on the appearance based visual front end of our audio-visual system. Section 3 is devoted to AV-ASR, with various audio-visual integration methods discussed within the framework of hidden Markov models [28]. The methods are grouped into the general categories of feature, decision, and hybrid fusion. Section 4 follows with the presentation of both a linear and a non-linear approach for bimodal based audio feature enhancement. A comparative experimental study of the discussed techniques is subsequently presented in Section 5, and the paper concludes with a short discussion.

2. The Visual Front End

A prerequisite for performing audio-visual ASR or enhancement is the successful extraction of visual features that are informative about the spoken utterance. Various possibilities exist for the visual front end design, and are briefly discussed below, followed by the particular implementation in our system.

2.1. Taxonomy and components of the visual front end

Visual speech features generally fit into one of the following three categories: *Appearance* based features, *shape* based ones, or combination of both [25]. The first assume that all video pixels within a *region-of-interest* (ROI) are informative about the spoken utterance. To allow speech classification, they consider mostly linear transforms of the ROI pixel values, resulting in



Figure 1: Face, facial part detection, and ROI extraction for an example video frame. Left-to-right: Original frame with eleven detected facial parts super-imposed; face-area enhanced frame; normalized ROI.

feature vectors of reduced dimensionality that contain most relevant speech information [23, 33, 34]. In contrast, shape based feature extraction assumes that most speechreading information is contained in the contours of the speaker’s lips, or more generally, of the face. Within this category belong geometric type features, such as mouth height, width, and area [18, 24], lip-contour Fourier descriptors [35], lip image moments [34], and statistical models of shape, such as active shape models [36], or other parameters of lip-tracking models [37]. Finally, features from both categories can be concatenated into a joint shape and appearance vector [38, 39], or a joint statistical model can be learned on such vectors, as is the case of the active appearance model, used in [26].

Clearly, a number of video pre-processing steps are required before the above mentioned visual feature extraction techniques can commence. One such step is face and facial part detection, that is needed to drive the ROI extraction (see also Fig.1). Face detection has attracted significant interest in the literature [40, 41], and it constitutes a difficult problem, especially in cases where the background, head pose, and lighting are varying. Of course, face detection is unnecessary if a properly head-mounted video camera is used to directly provide the ROI [42]. On the other hand, if shape-based visual features are to be extracted, the additional step of lip and possibly face shape estimation is also required.

2.2. Audio-visual features in our system

The visual front end employed in our system produces appearance based features and operates on full face video with no artificial face markings. As a result, both face detection and ROI extraction are required. In more detail, given the video of a spoken utterance, a two-stage statistical face tracking algorithm is first used to detect the speaker’s face and subsequently locate 26 facial features (eleven such features are depicted in Fig.1). At each stage, normalized face (or facial feature) candidate vectors are scored by a two-class Fisher discriminant and their projection residual onto an appropriately defined eigenspace [41]. The highest score candidates are retained as detected faces (or facial features). The algorithm requires training on a small number of manually annotated faces.

Tracking provides the mouth location, size, and orientation, which are then smoothed over time to improve robustness. Based on the resulting estimates, a 64×64 pixel ROI is obtained for every video frame. This contains the lower face around the speaker’s mouth, and is properly normalized to compensate for rotation, size, and lighting variations (see also Fig.1).

Subsequently, a two-dimensional, separable *discrete cosine transform* (DCT) is applied to the ROI, and the 100 highest-energy DCT coefficients are retained. To reduce dimensionality and improve discrimination among the speech classes, an *intra-frame linear discriminant analysis* (LDA) projection is applied, resulting in a 30-dimensional feature vector. This is followed by a *maximum likelihood linear transformation* (MLLT) [26], that improves maximum likelihood based statistical data mod-

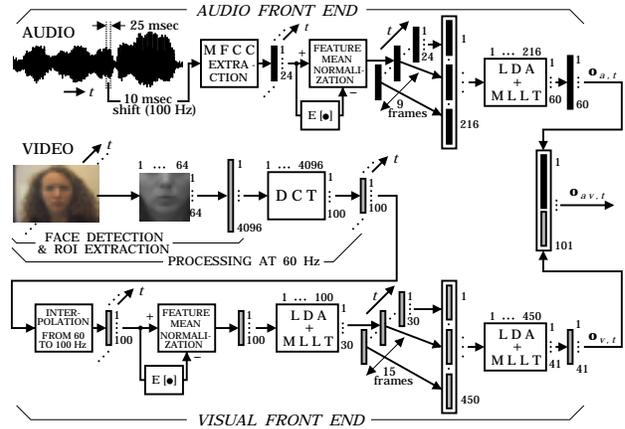


Figure 2: Block diagram of the front end for AV-ASR. The algorithm generates time-synchronous 60-dimensional audio feature vectors and 41-dimensional visual observations, both at a 100 Hz rate.

eling. To facilitate audio-visual fusion, linear interpolation is employed that synchronizes the features to the 100 Hz rate of their audio counterpart, whereas feature mean normalization is used to further compensate for lighting variations, providing the visual-only *static* features. Fifteen consecutive such features are then concatenated, and projected/rotated by means of an *inter-frame* LDA/MLLT combination, thus giving rise to *dynamic* visual features $\mathbf{o}_{v,t}$ of dimension $l_v = 41$ (see also Fig.2).

In addition to visual features, time-synchronous audio features are extracted at 100 Hz. First, 24 *mel-frequency cepstral coefficients* (MFCC) of the speech signal are computed over a sliding window of 25 msec, and are mean normalized to provide static features. Then, nine consecutive such features are concatenated and projected by means of LDA and MLLT onto dynamic audio features $\mathbf{o}_{a,t}$ of dimension $l_a = 60$.

3. Audio-Visual ASR

Audio-visual integration aims at combining the two available speech informative streams into a bimodal classifier with superior performance to both audio- and visual-only recognition. Various information fusion algorithms have been considered for AV-ASR, differing in their basic design, the speech classification technology used, as well as in the adopted terminology [25, 27, 43]. In this paper, we solely consider the traditional *hidden Markov model* (HMM) based approach for ASR [28], that employs acoustic-based speech classes and Gaussian mixture densities as the class-conditional probabilities of the feature observations of interest. Thus, a number of viable alternatives such as hybrid HMM/neural network [33, 44], or support vector machine [45] based ASR architectures, possibly using visual-based speech classes [45, 46], are not discussed.

We adopt a broad grouping of audio-visual integration techniques into *feature fusion* and *decision fusion* methods. The first are based on training a single classifier (i.e., of the same form as the audio- and visual-only ones) on the concatenated vector of audio and visual features, or on any appropriate transformation of it [24, 26, 43]. In contrast, decision fusion algorithms utilize the two single-modality (audio- and visual-only) classifier outputs to recognize audio-visual speech. Typically, this is achieved by linearly combining the class-conditional observation log-likelihoods of the two classifiers into a joint audio-visual score, using appropriate weights that capture the reliability of each single-modality data stream [39, 47]. In addition to

the above categories, there exist techniques that combine characteristics of both. Here, we consider one such *hybrid fusion* method (see also Fig.3). The presentation of all techniques initially assumes an “early” temporal level of audio-visual integration, namely at the HMM state. So-called “asynchronous” models of fusion are discussed at the end of the section.

3.1. Feature Fusion

Audio-visual feature fusion techniques include: Plain feature *concatenation* [24], feature weighting [43], both also known as *direct identification* fusion [43], hierarchical *discriminant* feature extraction [26], as well as the *dominant* and *motor* recording fusion [43]. The latter seek a data-to-data mapping of either the visual features into the audio space, or of both modality features to a new common space, followed by linear combination of the resulting features. In this paper, we briefly review two feature fusion methods.

Given time-synchronous audio and visual feature vectors, concatenative feature fusion considers

$$\mathbf{o}_{av,t} = [\mathbf{o}_{a,t}, \mathbf{o}_{v,t}] \in \mathbb{R}^{l_{av}}, \text{ where } l_{av} = l_a + l_v, \quad (1)$$

as the joint audio-visual observation of interest. A sequence of such features is assumed to be generated by a *single-stream* HMM, with class-conditional observation probabilities

$$P(\mathbf{o}_{s,t} | c) = \sum_{k=1}^{K_{s,c}} w_{s,c,k} \mathcal{N}_{l_s}(\mathbf{o}_{s,t}; \mathbf{m}_{s,c,k}, \mathbf{s}_{s,c,k}), \quad (2)$$

for all speech classes c , where $s = av$. In (2), $K_{s,c}$ mixture weights $w_{s,c,k}$ are positive and add to one, and $\mathcal{N}_l(\mathbf{o}; \mathbf{m}, \mathbf{s})$ denotes the l -variate normal distribution with mean \mathbf{m} and a diagonal covariance matrix \mathbf{s} .

In practice, l_{av} can be large, causing inadequate modeling in (2) due to the curse of dimensionality and insufficient data. Discriminant feature fusion aims to remedy this, by applying an LDA projection on the concatenated vector $\mathbf{o}_{av,t}$. Such projection results in a lower dimensional representation of (1), while seeking the best discrimination among the speech classes of interest. LDA can be followed by an MLLT rotation of the feature vector to improve statistical data modeling by means of (2), as in [26]. The transformed audio-visual feature vector is denoted by $\mathbf{o}_{d,t}$ in Fig.3, and in this work, it is designed to be of the same dimension as the audio observation, i.e., $l_d = l_a$.

Both concatenative and discriminant feature fusion are implementable in most existing ASR systems with minor changes, due to their use of single-stream HMMs. All required HMM parameters can be estimated using the *expectation-maximization* (EM) algorithm on available training data [28].

3.2. Decision Fusion

Although many feature fusion techniques result in improved ASR over audio-only performance [26], they cannot explicitly model the reliability of each modality, which in practice varies. The decision fusion framework, on the other hand, provides a mechanism for capturing these reliabilities, by borrowing from classifier combination theory [48]. For AV-ASR, most commonly, the audio- and visual-only classifiers are combined using a parallel architecture, adaptive combination weights, and class score level information. This approach derives the most likely speech class (or word sequence) by linearly combining the log-likelihoods of the two single-modality classifier decisions, using appropriate weights [24, 39, 47], and it is also known as the *separate identification* model [43, 46].

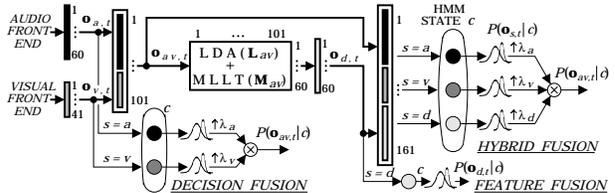


Figure 3: Representative techniques of the three fusion categories, considered in this paper for AV-ASR.

In the case where single-stream HMMs, with the same set of speech classes (states), are used for both audio- and visual-only classification, as in (2), this type of likelihood combination can be considered at a frame (HMM state) level, and modeled by means of the *multi-stream* HMM [49]. For a two-stream HMM, the state-dependent emission of the audio-visual observation vector $\mathbf{o}_{av,t}$ is governed (see also (1) and (2)) by [26, 39]

$$P(\mathbf{o}_{av,t} | c) = P(\mathbf{o}_{a,t} | c)^{\lambda_{a,c,t}} P(\mathbf{o}_{v,t} | c)^{\lambda_{v,c,t}}, \quad (3)$$

for all HMM states c . Notice that (3) implies a linear combination in the log-likelihood domain, but does not represent a probability distribution in general. In (3), $\lambda_{s,c,t}$ denote the stream exponents (weights), that are non-negative, and model stream reliability as a function of modality s , HMM state c , and utterance frame t . Typically, they are set to global, modality-only dependent values, λ_s , but some works also investigate their dependence on HMM state [50], or utterance frame [27].

Here, we assume global exponents, constrained to add up to one. These are estimated by simple grid search to minimize the word error rate on a held-out set. Alternatively, discriminative training can be used [47]. The remaining HMM parameters can be estimated separately for each stream using the EM algorithm, or jointly using (3) at the E-step. The latter scheme enforces state synchrony in training and is thus preferable.

3.3. Hybrid Fusion

Certain feature fusion techniques, for example discriminant fusion, outperform audio- and visual-only ASR [26]. It therefore seems natural to utilize $\mathbf{o}_{d,t}$ as a stream in multi-stream based decision integration (3), thus combining feature and decision fusion within the framework of the latter. In this paper, we consider two such hybrid approaches, by generalizing the two-stream HMM of (3) into

$$P(\mathbf{o}_{av,t} | c) = \prod_{s \in S} P(\mathbf{o}_{s,t} | c)^{\lambda_s}, \quad (4)$$

where $S = \{a, v, d\}$, or $S = \{a, d\}$. In the first case, we obtain a three-stream HMM, with the added stream of discriminant features $\mathbf{o}_{d,t}$. In the second case, we retain the two-stream HMM, however after replacing the less speech-informative visual stream with its superior discriminant audio-visual feature stream. As discussed above, the exponents are constrained by $\lambda_s \geq 0$ and $\sum_{s \in S} \lambda_s = 1$, whereas parameter estimation of the HMM components can be performed separately, or jointly. A schematic representation of hybrid fusion is depicted in Fig.3.

3.4. Audio-Visual Asynchrony in Fusion

In our presentation of decision and hybrid fusion, we have assumed the “early” temporal level of HMM states for combining the stream likelihoods of interest (see (3) and (4)). In ASR however, sequences of classes (HMM states or words) need to be

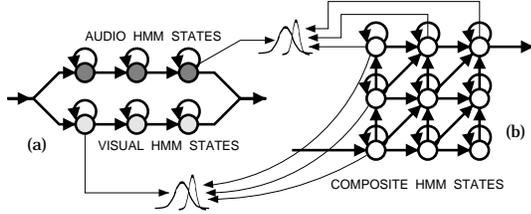


Figure 4: (a) Phone-synchronous (state-asynchronous) two-stream HMM with three states per phone and modality. (b) Its equivalent product (composite) HMM. The single-stream emission probabilities are tied for states along the same row (column) to the corresponding audio (visual) state probabilities of form (2), according to (5).

estimated, therefore coarser levels for combining stream likelihoods can also be envisioned. One such “late” level of integration can be the utterance end, where typically a number of N -best hypotheses (or all vocabulary words, in case of isolated word recognition) are rescored by the stream log-likelihoods, independently computed over the entire utterance. An example of late fusion is the discriminative model combination technique applied for AV-ASR in [26]. Alternatively, the phone, syllable, or word boundary can provide an “intermediate” level of integration. Such a scheme is typically implemented by means of the *product*, or *coupled* HMM, as discussed next. Notice that both approaches permit asynchrony between the HMM state sequences of the streams of interest, thus providing the means to model the actual audio and visual signal asynchrony, observed in practice to be up to the order of 100 msec [33].

The product HMM [26, 39, 51] is a generalization of the state-synchronous multi-stream HMM (4) that combines the stream log-likelihoods at an intermediate level, here assumed to be the phone. The resulting phone-synchronous product HMM allows its single-stream HMM components to be in asynchrony within each phone, forcing their synchrony at the phone boundaries instead. It consists of *composite* states $\mathbf{c} = \{c_s, s \in \mathcal{S}\}$ with emission scores similar to (4), namely

$$P(\mathbf{o}_{av,t} | \mathbf{c}) = \prod_{s \in \mathcal{S}} P(\mathbf{o}_{s,t} | c_s)^{\lambda_s}. \quad (5)$$

An example of such a model is depicted in Fig.4, for the typical case of one audio and one visual stream, and three states per phone and stream. Notice that in (5), the stream components correspond to the emission probabilities of certain single-stream states, tied as demonstrated in Fig.4. Therefore, compared to its corresponding state-synchronous multi-stream HMM, the product HMM utilizes the same number of mixture weight, mean, and variance parameters. On the other hand, the number of transition probabilities between its composite states is larger. Such probabilities between states \mathbf{c}' and \mathbf{c}'' are often factored as $P(\mathbf{c}' | \mathbf{c}'') = \prod_{s \in \mathcal{S}} P(c'_s | c''_s)$, in which case the resulting model is referred to as the coupled HMM [52, 53].

4. Bimodal Enhancement of Audio Features

In addition to improving ASR, the visual modality has been investigated as a means of noisy audio enhancement. For example, Girin, et. al., in [29], propose estimating clean audio features (linear prediction model coefficients, and subsequently the clean audio signal) from visual speech information, whereas in [30] they consider estimating such features from audio-visual speech, when the audio channel is corrupted by noise. Such an approach proves feasible, due to the fact that audio and visible speech are produced by the same oral-facial cavity, and hence

are correlated. Indeed, audio feature estimation from visual input has also been demonstrated in [7, 8].

Clearly, audio-visual ASR and audio-visual speech enhancement differ in their aims and methodologies; however, one expects that the latter would also lead to improved recognition performance over the use of a noisy audio-only based ASR system. Furthermore, enhancing the noisy audio features could enable the use of clean audio statistical models for ASR over a wide variety of noisy environments, thus avoiding noise-dependent statistical model training and storage. Therefore, it is of interest to study the effects of audio-visual speech enhancement to ASR and to compare the resulting system performance to traditional audio-visual ASR.

In this section, we summarize our work on two techniques for enhancing noisy audio features based on bimodal data. The first method is linear [31], and is based on the algorithm reported in [30]. It enhances noisy audio features by means of a linear filter (transform), which is applied on the concatenated vector of noisy audio and visual features. Similarly to that work, the filter is obtained by *mean square error* (MSE) estimation of the clean audio feature training data. Since we are interested in ASR, we do not consider the problem of obtaining enhanced speech from the enhanced audio features. So, instead of using linear prediction coefficients, as in [29, 30], we use the MFCC-based audio features of our system. The second technique is introduced in [32] and is non-linear. It constitutes an extension of an audio enhancement method, known as *codebook dependent cepstral normalization* (CDCN) [21, 22], with the visual modality utilized to improve the estimation of the correction term applied to the noisy audio features. The resulting method is referred to as *audio-visual codebook dependent cepstral normalization* (AVCDCN). An alternative non-linear enhancement technique based on neural networks is reported in [30].

4.1. Linear Bimodal Enhancement of Audio Features

In addition to speech information, the audio feature vector $\mathbf{o}_{a,t}$ extracted by the front end of Section 2 captures environment noise. We hope to remove such interference and produce *enhanced* audio features, that we denote by $\mathbf{o}_{a,t}^{(E)} \in \mathbb{R}^{l_a}$, using the joint audio-visual speech information captured in vector (1). The resulting enhanced audio features can then be supplied to an ASR system, hopefully yielding improved recognition over the use of noisy observations, $\mathbf{o}_{a,t}$.

As in [30, 31], we seek to obtain enhanced audio observations $\mathbf{o}_{a,t}^{(E)}$ as a linear transformation of the joint audio-visual feature vector $\mathbf{o}_{av,t}$, namely as

$$\mathbf{o}_{a,t}^{(E)} = \mathbf{o}_{av,t} \mathbf{P}_{av}^{(E)}, \quad (6)$$

where matrix $\mathbf{P}_{av}^{(E)} = [\mathbf{p}_{av,1}^\top, \mathbf{p}_{av,2}^\top, \dots, \mathbf{p}_{av,l_a}^\top]$ is of dimension $l_{av} \times l_a$, its columns consisting of l_{av} -dimensional vectors $\mathbf{p}_{av,i}^\top$, for $i = 1, \dots, l_a$. To estimate matrix $\mathbf{P}_{av}^{(E)}$, we assume that in addition to (1), *clean* audio feature vectors, denoted by $\mathbf{o}_{a,t}^{(C)}$, are available for a number of instants t in a training set, \mathcal{T} . We then seek to estimate the enhancement matrix in (6), such that $\mathbf{o}_{a,t}^{(E)} \approx \mathbf{o}_{a,t}^{(C)}$ over set \mathcal{T} , in the Euclidean distance sense. Due to (6), this is equivalent to solving l_a MSE estimations

$$\mathbf{p}_{av,i} = \arg \min_{\mathbf{p}} \sum_{t \in \mathcal{T}} [o_{a,t,i}^{(C)} - \langle \mathbf{p}, \mathbf{o}_{av,t} \rangle]^2, \quad (7)$$

for $i = 1, \dots, l_a$, i.e., one per column of matrix $\mathbf{P}_{av}^{(E)}$. Equations

(7) result to l_a systems of the Yule-Walker equations

$$\sum_{j=1}^{l_{av}} \left[\sum_{t \in \mathcal{T}} o_{av,t,j} o_{av,t,k} \right] p_{av,i,j} = \sum_{t \in \mathcal{T}} o_{a,t,i}^{(C)} o_{av,t,k}, \quad (8)$$

for $k = 1, \dots, l_{av}$, where $p_{av,i,j}$ denotes the j -th element of vector $\mathbf{p}_{av,i}$, and $o_{s,t,i}$ the i -th element of feature vector $\mathbf{o}_{s,t}$. Gauss-Jordan elimination can be used to solve (8).

4.2. Audio-Visual CDCN for Audio Feature Enhancement

As demonstrated in Section 5, the ASR performance of the above linear enhancement approach is mediocre: Although recognition improves compared to audio-only ASR using the noisy acoustic observations, performance remains inferior to AV-ASR by means of discriminant feature fusion, for example. To break this barrier, non-linear techniques are required. One such method is AVCDCN, introduced in [32]. The technique is inspired from CDCN [21, 22], a popular audio-only non-linear enhancement approach. In CDCN, the non-linear effect of the noise on the clean speech features is approximated with a piece-wise constant function. AVCDCN is a multi-sensor extension of CDCN that integrates the use of audio and visual features. Our experiments show that the use of visual information in AVCDCN allows significant performance gains over CDCN, as well as over feature fusion AV-ASR, when using HMMs trained in the clean acoustic environment.

The CDCN technique seeks to compute enhanced audio features $\mathbf{o}_{a,t}^{(E)}$ as the expected value of $\mathbf{o}_{a,t}^{(C)}$ given the observed noisy audio features $\mathbf{o}_{a,t}$ [21, 22]. Namely,

$$\mathbf{o}_{a,t}^{(E)} = \int_{\mathbf{o}_a^{(C)}} \mathbf{o}_a^{(C)} P(\mathbf{o}_a^{(C)} | \mathbf{o}_{a,t}) d\mathbf{o}_a^{(C)}. \quad (9)$$

Using the fact that $\mathbf{o}_{a,t}^{(C)} = \mathbf{o}_{a,t} - f(\mathbf{o}_{a,t}^{(C)}, \mathbf{n}_t)$, where $f(\bullet, \bullet)$ is a non-linear function of the clean audio and corrupting noise \mathbf{n}_t [21], (9) becomes

$$\mathbf{o}_{a,t}^{(E)} = \mathbf{o}_{a,t} - \int_{\mathbf{o}_a^{(C)}} f(\mathbf{o}_a^{(C)}, \mathbf{n}_t) P(\mathbf{o}_a^{(C)} | \mathbf{o}_{a,t}) d\mathbf{o}_a^{(C)}. \quad (10)$$

The novelty in AVCDCN is to use visual modality features $\mathbf{o}_{v,t}$, in addition to the traditional acoustic vector $\mathbf{o}_{a,t}$, in order to more accurately estimate the correction term applied to the latter in (10). Namely,

$$\mathbf{o}_{a,t}^{(E)} = \mathbf{o}_{a,t} - \int_{\mathbf{o}_a^{(C)}} f(\mathbf{o}_a^{(C)}, \mathbf{n}_t) P(\mathbf{o}_a^{(C)} | \mathbf{o}_{av,t}) d\mathbf{o}_a^{(C)}. \quad (11)$$

For lack of knowledge of $\mathbf{o}_a^{(C)}$ and \mathbf{n}_t , we approximate (10) and (11) with a sum over a pre-defined codebook of audio compensation terms $\{f_{a,k}\}_{k=1}^K$, computed as discussed next. Thus, AVCDCN yields

$$\mathbf{o}_{a,t}^{(E)} = \mathbf{o}_{a,t} - \sum_{k=1}^K f_{a,k} P(k | \mathbf{o}_{av,t}), \quad (12)$$

whereas its audio-only counterpart, CDCN, is defined by:

$$\mathbf{o}_{a,t}^{(E)} = \mathbf{o}_{a,t} - \sum_{k=1}^K f_{a,k} P(k | \mathbf{o}_{a,t}). \quad (13)$$

Note that both AVCDCN and CDCN use identical audio compensation codewords, however AVCDCN takes advantage of



Figure 5: Example video frames from the corpus considered in this paper for audio-visual recognition and enhancement of large-vocabulary continuous speech (upper row) and connected digits (lower row).

the visual information to estimate the posterior distribution of the codewords, as follows.

The codebook posterior distribution $\{P(k | \mathbf{o}_{av,t})\}_{k=1}^K$ is computed by assuming that the *probability density function* (PDF) of $\mathbf{o}_{av,t}$ is a mixture of Gaussians with priors, means and covariances $(w_k, \mathbf{m}_k, \mathbf{S}_k)_{k=1}^K$, and thus by Bayes' rule,

$$P(k | \mathbf{o}_{av,t}) = \frac{w_k \mathcal{N}_{l_{av}}(\mathbf{o}_{av,t}; \mathbf{m}_k, \mathbf{S}_k)}{\sum_{k'=1}^K w_{k'} \mathcal{N}_{l_{av}}(\mathbf{o}_{av,t}; \mathbf{m}_{k'}, \mathbf{S}_{k'})}. \quad (14)$$

In our experiments, both the codebook of audio compensations and the PDF parameters of the noisy audio-visual features are estimated from “stereo” training data $\{(\mathbf{o}_{av,t}, \mathbf{o}_{a,t}^{(C)})\}_{t \in \mathcal{T}}$.

The audio compensations are computed by minimizing the MSE between $\mathbf{o}_{a,t}^{(C)}$ and $\mathbf{o}_{a,t}$ over set \mathcal{T} , i.e.,

$$f_{a,k} = \frac{\sum_{t \in \mathcal{T}} (\mathbf{o}_{a,t} - \mathbf{o}_{a,t}^{(C)}) P(k | \mathbf{o}_{a,t}^{(C)})}{\sum_{t \in \mathcal{T}} P(k | \mathbf{o}_{a,t}^{(C)})}. \quad (15)$$

Maximum likelihood estimates of the means and covariances of the noisy audio-visual features are computed as (assuming equal priors):

$$\begin{aligned} \mathbf{m}_k &= \frac{\sum_{t \in \mathcal{T}} \mathbf{o}_{av,t} P(k | \mathbf{o}_{a,t}^{(C)})}{\sum_{t \in \mathcal{T}} P(k | \mathbf{o}_{a,t}^{(C)})} \\ \mathbf{S}_k &= \frac{\sum_{t \in \mathcal{T}} (\mathbf{o}_{av,t} - \mathbf{m}_k)^\top (\mathbf{o}_{av,t} - \mathbf{m}_k) P(k | \mathbf{o}_{a,t}^{(C)})}{\sum_{t \in \mathcal{T}} P(k | \mathbf{o}_{a,t}^{(C)})}. \end{aligned} \quad (16)$$

The posteriors $P(k | \mathbf{o}_{a,t}^{(C)})$ are computed by assuming that the PDF of the clean audio features $\mathbf{o}_{a,t}^{(C)}$ is a mixture of Gaussians with equal priors, and with means and covariances for which maximum likelihood estimates are computed with a standard EM algorithm on the clean audio training data. In our CDCN baseline, the means and covariances of the noisy audio features are computed by replacing $\mathbf{o}_{av,t}$ by $\mathbf{o}_{a,t}$ in (16).

5. Experiments

So far, we have discussed a number of techniques for audio-visual ASR and bimodal enhancement of audio features. We now proceed to report speech recognition experiments on a suitable audio-visual database using these algorithms. We first discuss the corpus, briefly introduce the experimental paradigm adopted, with a subsequent detailed presentation of our results.

5.1. The Audio-Visual Database

In contrast to the abundance of audio-only corpora, there exist only a few databases suitable for bimodal ASR research. Most of them are rather small, contain few subjects, and address simple recognition tasks, such as small-vocabulary ASR of isolated or connected words [9, 25].

Task	Set	Utter.	Dur.	Sub.
LVCSR	Train	17111	34:55	239
	Check	2277	4:47	25
	Test	670	2:29	26
DIGIT	Train	5490	8:01	50
	Check	670	0:58	50
	Test	529	0:46	50

Table 1: The corpus partitioning into training, check (held-out), and test sets, used in the large-vocabulary continuous speech (LVCSR) and connected digit (DIGIT) recognition experiments of Section 5 (number of utterances, duration (in hours), and number of subjects are shown).

To help bridge the growing gap between audio-only and AV-ASR corpora, we have collected the IBM ViaVoiceTM audio-visual database, a large corpus suitable for speaker-independent audio-visual LVCSR. The corpus consists of full-face frontal video and audio of 290 subjects (see also Fig.5), uttering ViaVoiceTM training scripts, i.e., continuous read speech with mostly verbalized punctuation, dictation style (a 10.4k-word vocabulary is used). The data are collected using a teleprompter in a quiet studio environment. In more detail, the video is of a 704 × 480 pixel size, interlaced, captured in color at a rate of 30 Hz (60 fields per second are available at a resolution of 240 lines), and it is MPEG2 encoded at the relatively high compression ratio of about 50:1. High quality, wideband audio is synchronously recorded at a rate of 16 kHz and a *signal-to-noise ratio* (SNR) of 19.5 dB. In addition to the LVCSR data, we have collected a smaller, 50-subject set, containing utterances of 7- and 10-digit connected strings (both “zero” and “oh” are used). This is recorded under the same conditions as the LVCSR set, and is referred to in this work as the DIGIT corpus.

5.2. The Experimental Paradigm

For all single-stream HMM based recognition tasks, we use 3-state, left-to-right phone HMMs, with context-dependent sub-phonetic classes (states). These classes are obtained by means of decision trees that cluster contexts spanning up to 5 phones to each side of the current phone, in order to better model co-articulation and improve ASR performance. Both DIGIT and LVCSR decision trees are estimated using the clean audio of the corresponding database training set, by bootstrapping on a previously developed audio-only HMM (and its corresponding front end), which provides data class labels by forced alignment. Subsequently, K -means clustering is used to estimate audio-only HMMs, that correspond to the newly developed trees. It is by bootstrapping on these models, that the parameters of all HMMs considered in this paper are estimated (on their required front ends). The total number of the resulting context-dependent HMM states are 159 for the DIGIT task (corresponding to 22 phones) and approximately 2.8k for LVCSR (for 52 phones). All single-stream HMMs have identical number of Gaussian mixture components, namely about 3.2k and 47k for the DIGIT and LVCSR tasks, respectively.

Once decision trees and initial DIGIT and LVCSR audio HMMs are developed, we proceed to estimate the parameters of single-stream HMMs that model visual-only, as well as audio-only and audio-visual feature sequences at a number of audio channel conditions. Both the original clean database audio at approximately 19.5 dB SNR, as well as noisy conditions, where speech babble noise is artificially added at various SNRs, are considered. We use three EM algorithm iterations for training, with the E-step of the first iteration employing the initial audio-only HMM (for bootstrapping). Appropriate single-

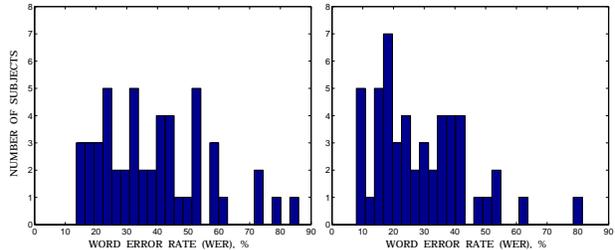


Figure 6: Visual-only ASR on the DIGIT test set, depicted as the word-error-rate histogram of the 50 subjects, achieved by visual-only HMMs trained in a speaker-independent (left) or multi-speaker fashion (right).

stream HMMs are also joined to form the decision and hybrid fusion models of Section 3, i.e., (3), (4), and (5), with the stream exponents set to global values, estimated on the held-out sets of Table 1. Joint stream HMM training is also considered.

For AV-ASR, all results are reported on recognition of matched test data (same SNR as in training). For the DIGIT task, decoding is based on a simple digit-word loop grammar (with unknown string length), whereas for LVCSR, a trigram language model is used. In both cases, a two-stage stack decoding algorithm is employed. LVCSR results are speaker-independent, whereas DIGIT recognition is multi-speaker.

For the bimodal enhancement of audio features, stereo pair data consisting of noisy audio-visual and clean audio observations are available on the training sets of Table 1. For the linear approach of Section 4.1, the regression matrix is computed on the SNR of interest, as in (8), and applied to bimodal input. The recognition performance of the resulting enhanced audio features is compared to both audio-only, as well as discriminant feature based ASR. For (AV)CDCN, computation of the PDF characterizing the clean speech in the audio channel of the training set is required. A set of audio compensation codewords and the PDF characterizing the noisy audio-visual speech (or the noisy audio-only speech, in the case of audio-only CDCN) are then estimated, for each SNR condition, according to (15) and (16). The PDFs of the noisy speech are estimated on the features output by the audio-visual LDA/MLLT transform for AVCDCN and on the features output by the audio LDA/MLLT transform for CDCN. When decoding the test set, the acoustic features are enhanced with either AVCDCN or CDCN according to (12) or (13), where the posterior probabilities are computed with the features output by the LDA/MLLT transforms and the PDFs of noisy speech matching the SNR level under consideration. The AVCDCN and CDCN enhancement strategies are evaluated for various sizes of codebooks across a number of SNR levels, using both audio and audio-visual ASR. This is benchmarked against audio and audio-visual ASR without enhancement. Furthermore, we report on recognition experiments where the LDA/MLLT transforms producing the features sent to the decoder and the HMMs used by the decoder are either trained on the clean training data or re-trained on the enhanced noisy training data matching the SNR level under consideration.

5.3. Visual-Only Recognition

Compared to traditional acoustic ASR in the clean environment, recognition on basis of visual-only information performs very poorly. For example, the visual-only *word error rate* (WER) is 93.5% on the LVCSR task and 23.6% on the DIGIT one. The results improve after per-speaker adaptation by maximum likelihood linear regression to 82.5% and 16.8%, respectively. Such performance varies significantly across subjects, as demon-

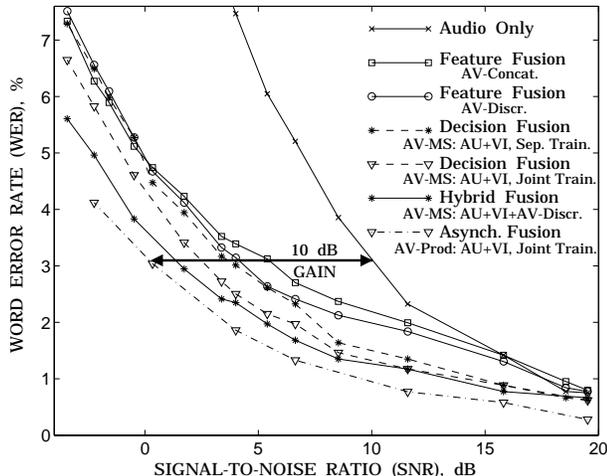


Figure 7: Audio-only and audio-visual ASR on the DIGIT test set using the integration strategies of Section 3. In all cases, WER, %, is depicted vs. audio channel SNR. The effective SNR gain using the product HMM is also shown, reported with reference to the audio-only WER at 10 dB. All HMMs are trained in matched noise conditions.

strated in Fig.6. There, a WER histogram of the 50 DIGIT dataset subjects is depicted, when using speaker-independent or multi-speaker visual-only HMMs. Clearly, visual features do provide speech information, albeit very weak. It is of course the combination with their audio counterpart that is of interest, as demonstrated next.

5.4. Audio-Visual ASR Experiments

We now proceed to investigate the visual feature benefit to ASR. For both LVCSR and connected-digit recognition, we consider acoustic conditions at a wide range of SNRs, as discussed in Section 5.1, and we compare the fusion strategies of Section 3 in terms of their resulting *effective SNR gain* in ASR. We measure this gain with reference to the audio-only WER at 10 dB, by considering the SNR value where the audio-visual WER equals the reference audio-only WER.

The performance of all integration algorithms on the DIGIT set is summarized in Fig.7. In more detail, we first compare AV-ASR by means of the two feature fusion methods of Section 3.1. As it becomes clear from Fig.7, both concatenative and discriminative feature fusion significantly improve ASR performance at low SNRs, with the latter being somewhat superior, yielding an approximate 6 dB of effective SNR gain. For example, at -2.2 dB SNR, discriminant fusion based AV-ASR results in a 6.3% WER, representing a vast improvement over the audio-only WER of 19.8%. Notice however that feature fusion fails to alter performance at the high end of the SNR range considered. On the other hand, decision based audio-visual integration, by means of the state-synchronous two-stream HMM discussed in Section 3.2, consistently improves performance at all SNRs. In particular, joint stream training of the model is clearly preferable, outperforming separate stream training and discriminant feature fusion, and yielding a 7.5 dB effective SNR gain. Further improvements (9 dB) can be obtained by using the hybrid fusion approach of Section 3.3 that utilizes the discriminant audio-visual features as an additional stream within a three-stream HMM. Finally, introducing state asynchrony in decision fusion results in further gains. A jointly trained product HMM achieves approximately a 10 dB SNR gain, thus exhibiting at 0 dB the performance of audio-only ASR at the much cleaner

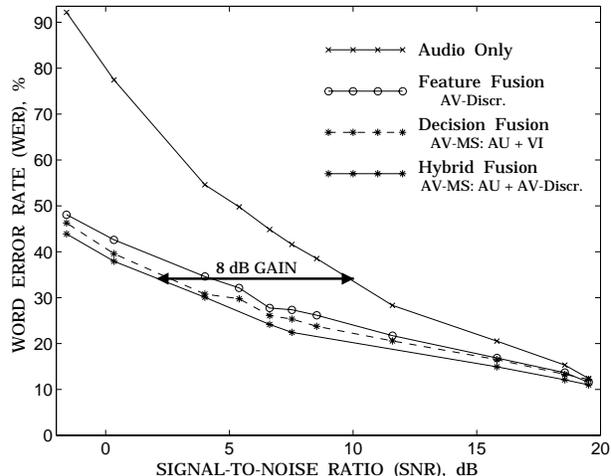


Figure 8: Audio-only and audio-visual WER, %, on the LVCSR test set using discriminant feature fusion, as well as two-stream HMMs for decision and hybrid fusion, for various SNR levels, matched to training.

acoustic environment of 10 dB. Notice that at -2.2 dB SNR, the product HMM yields a 4.1% WER, which corresponds to a 35% improvement over discriminant feature fusion and 79% over audio-only ASR. But even more remarkably, for the original database audio at 19.5 dB, the best audio-visual WER is 0.28%, which represents a 63% WER reduction over the audio-only WER of 0.75% (see also Fig.7). A large percentage of this gain is due to the joint estimation of all product HMM parameters with appropriate tying, since the composition of a product HMM by separately trained single-stream models achieves an inferior 0.40% WER.

For LVCSR, the performance of a number of the presented fusion techniques is summarized in Fig.8. Similarly to the results on the DIGIT set, hybrid fusion outperforms decision based integration, which in turn is superior to discriminant feature fusion, as well as audio-only ASR. For simplicity, a two-stream HMM is considered in hybrid fusion, where audio-visual discriminant features are used in place of the visual stream. The resulting system achieves approximately an 8 dB effective SNR gain over audio-only ASR at 10 dB.

5.5. Bimodal Enhancement Experiments

Following the demonstration of the visual modality benefit to ASR, we investigate its usefulness to the enhancement of noisy audio features. A simple way to quantify this is by benchmarking the ASR performance of the resulting enhanced features against the bimodal ASR results reported above.

We first consider the linear approach of Section 4.1. Fig.9 demonstrates that the enhanced audio features significantly outperform noisy audio-only ASR for matched training and testing, however fail to reach the performance of the system that discriminatively combines the audio and visual features. Thus, the investigated audio enhancement technique does not capture the full benefit of the visual modality to ASR. This observation holds for both DIGIT and LVCSR tasks.

Next we consider the non-linear enhancement technique of Section 4.2. Fig.10 shows how the WER obtained with AVDCN in an audio-only ASR scheme decreases consistently at all SNR values when the size of the codebook is increased from $K = 2$ to 128 codewords. On the same figure are also plotted the WERs obtained in an audio-only and audio-visual ASR scheme without feature enhancement. AVDCN outperforms

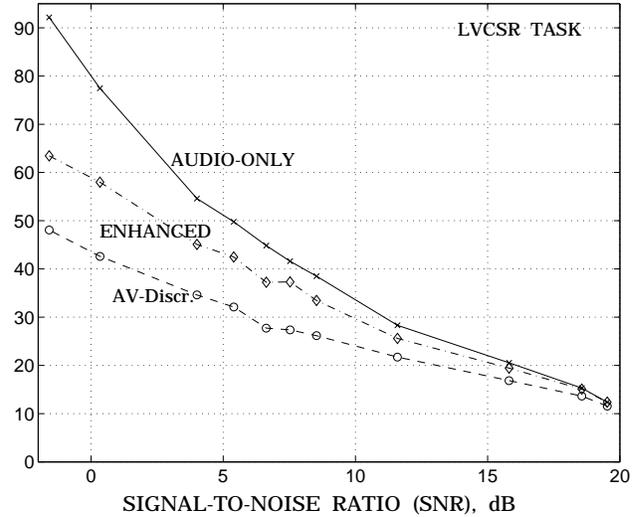
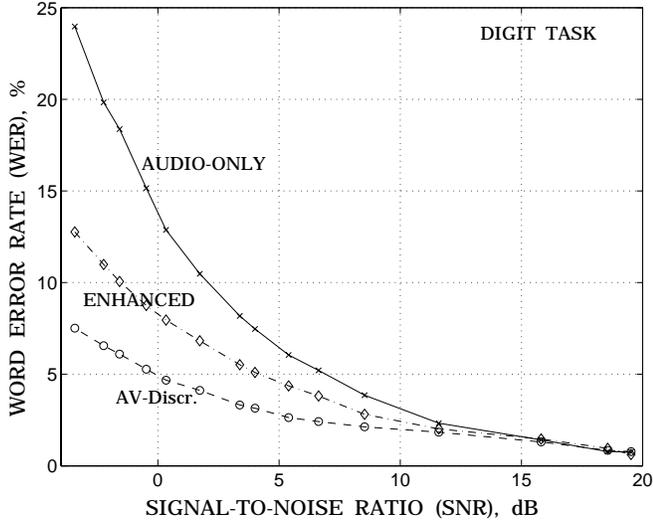


Figure 9: Test set WER (%) for noisy audio-only, audio-visually linearly enhanced audio, and discriminant audio-visual features, depicted against the audio channel SNR for connected digit ASR (left) and LVCSR (right).

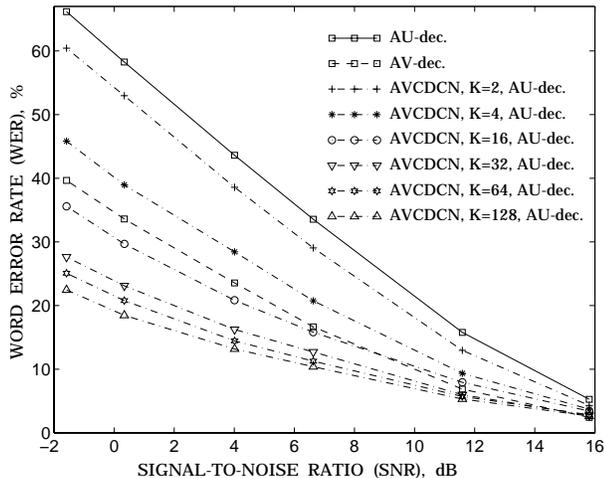


Figure 10: Audio-only ASR using AVCDCN-enhanced features for various codebook sizes. WERs are plotted against the audio-channel SNR. For comparison, the performance of audio-only and audio-visual ASR (using discriminant feature fusion) is also depicted. All HMMs are trained on clean data.

the audio-only ASR scheme regardless of codebook size, and most interestingly, also discriminant feature fusion for codebooks of size 16 and higher.

Fig.11 compares AVCDCN and CDCN in both audio and audio-visual ASR schemes for a codebook of size 128. Fig.11a shows WERs obtained with the recognition systems trained on the original clean training data. Fig.11b shows WERs obtained with the recognition systems retrained on: (i) the noisy training data matching the SNR level under consideration when no enhancement is used, (ii) the CDCN or AVCDCN-enhanced noisy training data matching the SNR level under consideration when either CDCN or AVCDCN is used. When systems are not retrained (Fig.11a), AVCDCN performs significantly better than CDCN in both the audio and audio-visual ASR schemes. Also, the performance gains obtained with AVCDCN and with audio-visual ASR add up, since AVCDCN combined with audio-visual ASR significantly outperforms both audio-visual ASR and AVCDCN combined with

audio-only ASR. Retraining the systems (Fig.11b) improves the performances of all strategies. AVCDCN is still better than CDCN in the audio ASR scheme. Besides, AVCDCN with audio-visual ASR outperforms AVCDCN with audio-only ASR. On the other hand, the performances of audio-visual ASR without enhancement, with CDCN and with AVCDCN become very similar.

6. Summary and Discussion

In this paper, we provided an overview of a number of techniques necessary in the automatic recognition of audio-visual speech, as well as the enhancement of noisy audio features on basis of audio-visual observations. We first discussed the visual front end that captures the speech information present in the video signal, and is shared in both.

For AV-ASR, we presented a number of fusion techniques, based on the popular hidden Markov model framework. We covered methods that integrate speech information at the feature or the classification score level, and presented a hybrid fusion algorithm that combines the benefits of both approaches. In addition, we discussed asynchrony modeling in audio-visual fusion, and we argued for the joint training of all properly tied parameters of the resulting model. The best technique, utilizing the product hidden Markov model, resulted in an effective SNR gain of 10 dB for connected-digit recognition. The best achieved gain on the large-vocabulary task was somewhat inferior, reaching approximately 8 dB.

We subsequently investigated the effects on ASR of enhancing noisy audio features by means of audio-visual speech data. We first considered enhancement performed by linear filters applied on concatenated audio-visual feature vectors. The method resulted to large improvements in ASR over the use of the original noisy audio features for both small- and large-vocabulary recognition. Compared however to audio-visual discriminant feature fusion, this enhancement approach fared significantly worse. We then discussed the generalization of CDCN, a non-linear audio-only based enhancement technique, to benefit from the availability of the visual channel. The resulting AVCDCN method was shown to provide significant performance gains over CDCN in both audio and audio-visual ASR schemes.

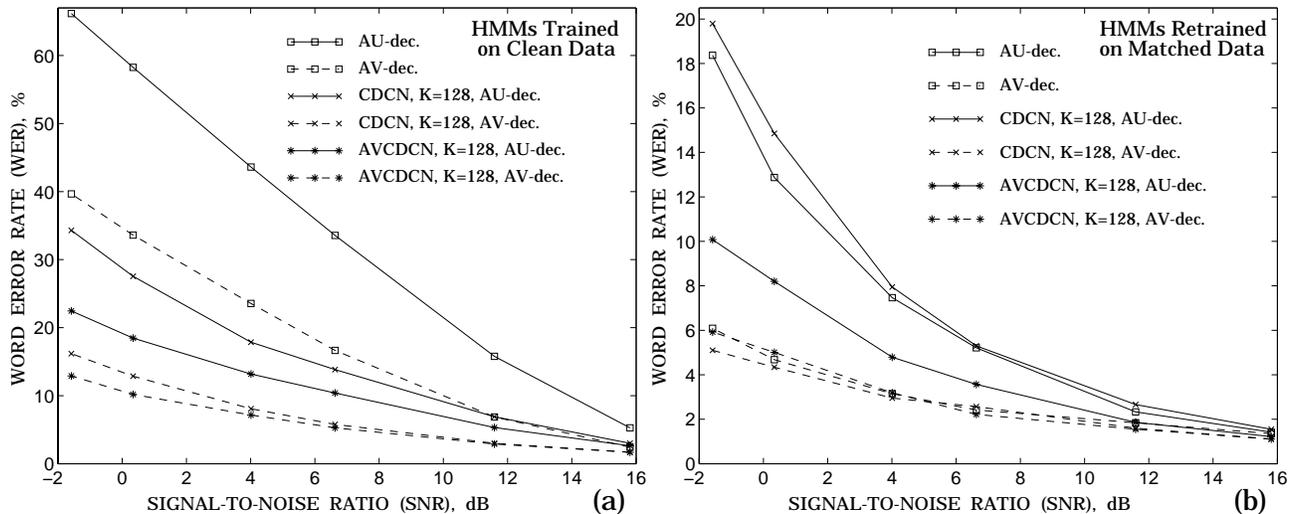


Figure 11: Audio-only and audio-visual ASR (by means of discriminant feature fusion) on noisy features, CDCN or AVDCN-enhanced features using: (a) HMMs trained on clean data, (b) HMMs trained on the noisy or enhanced features matching the SNR level under consideration.

The paper clearly demonstrates that over the past twenty years, much progress has been accomplished in capturing and integrating visual information into the speech recognition process. However, the visual modality has yet to become utilized in mainstream ASR systems. This is due to the fact that issues of both practical and research nature remain challenging. On the practical side of things, the high requirements in the captured video frame rate and size, necessary for extracting visual speech information that is capable of enhancing ASR performance, place increased demands on cost, storage, and computer processing. In addition, the lack of common, large audio-visual corpora that address a wide variety of ASR tasks, conditions, and environments, hinders development of audio-visual systems suitable for use in particular applications.

On the research side, key issues in the design of audio-visual recognition systems remain open and subject to more investigation. In the visual front end, for example, face, facial feature, and face shape tracking, robust to unconstrained speaker, pose, lighting, and environment variation constitutes a challenging problem. When combining audio and visual information, a number of issues relevant to decision fusion require further study, such as the optimal level of integrating the audio and visual log-likelihoods, the optimal function for this integration, and the robust modeling of channel reliability. When utilizing the visual modality for improved audio feature enhancement, investigation of AVDCN performance for unseen during training noise types and levels is also of interest. Further research on these issues is clearly warranted, and it is expected to lead to improving the value of audio-visual speech in the design of robust and natural human-computer interaction systems.

7. Acknowledgements

The authors would like to thank Guillaume Gravier, Roland Goecke, and Andrew W. Senior for contributing to this work.

8. References

- [1] D. G. Stork and M. E. Hennecke, Eds., *Speechreading by Humans and Machines*. Berlin, Germany: Springer, 1996.
- [2] R. Campbell, B. Dodd, and D. Burnham, Eds., *Hearing by Eye II*. Hove, United Kingdom: Psychology Press Ltd. Publishers, 1998.
- [3] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoustical Society America*, vol. 26, pp. 212–215, 1954.
- [4] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
- [5] A. Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading*, R. Campbell and B. Dodd, Eds. London, United Kingdom: Lawrence Erlbaum Associates, 1987, pp. 3–51.
- [6] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration," *American Scientist*, vol. 86, pp. 236–244, 1998.
- [7] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Commun.*, vol. 26, pp. 23–43, 1998.
- [8] J. P. Barker and F. Berthommier, "Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models," in *Proc. Conf. Audio-Visual Speech Processing*, Santa Cruz, CA, Aug. 7–9, 1999, pp. 112–117.
- [9] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol. 4, pp. 23–37, Mar. 2002.
- [10] X. Zhang, C. C. Broun, R. M. Mersereau, and M. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1228–1247, Nov. 2002.
- [11] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1154–1164, Nov. 2002.
- [12] P. De Cuetos, C. Neti, and A. Senior, "Audio-visual intent to speak detection for human computer interaction," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Istanbul, Turkey, June 5–9, 2000, pp. 1325–1328.
- [13] D. Sodooyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources: A new approach exploiting the audio-visual coherence of speech stimuli," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1165–1173, Nov. 2002.
- [14] E. Foucher, L. Girin, and G. Feng, "Audiovisual speech coder: Using vector quantization to exploit the audio/video correlation," in *Proc. Conf. Audio-Visual Speech Processing*, Terrigal, Australia, Dec. 4–6, 1998, pp. 67–71.
- [15] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. Wong, "Integration of multimodal features for video scene classification based on HMM," in *Proc. Works. Multimedia Signal Processing*, Copenhagen, Denmark, Sept. 13–15, 1999, pp. 53–58.

- [16] M. M. Cohen and D. W. Massaro, "What can visual speech synthesis tell visual speech recognition?" in *Proc. Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, 1994.
- [17] E. Cosatto and H. P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Trans. Multimedia*, vol. 2, pp. 152–163, Sept. 2000.
- [18] E. D. Petajan, "Automatic lipreading to enhance speech recognition," in *Proc. Global Telecomm. Conf.*, Atlanta, GA, 1984, pp. 265–272.
- [19] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, pp. 1–15, 1997.
- [20] R. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "Signal processing for robust speech recognition," in *Automatic Speech and Speaker Recognition. Advanced Topics*, C.-H. Lee, F. K. Soong, and Y. Ohshima, Eds. Norwell, MA: Kluwer Academic Pub., 1997, ch. 15, pp. 357–384.
- [21] A. Acero and R. Stern, "Environmental robustness in automatic speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Albuquerque, NM, 1990, pp. 849–852.
- [22] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo data," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, UT, May 7–11, 2001, pp. 301–304.
- [23] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading," in *Proc. Int. Conf. Spoken Lang. Processing*, Yokohama, Japan, Sept. 18–22, 1994, pp. 547–550.
- [24] 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. Berlin, Germany: Springer, 1996, pp. 461–471.
- [25] 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. Berlin, Germany: Springer, 1996, pp. 331–349.
- [26] C. Neti, G. Potamianos, J. Luetttin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, Final Workshop 2000 Report, Oct. 2000.
- [27] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, 2003, to appear.
- [28] J. R. Deller, Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Englewood Cliffs, NJ: Macmillan Publishing Company, 1993.
- [29] L. Girin, G. Feng, and J.-L. Schwartz, "Noisy speech enhancement with filters estimated from the speaker's lips," in *Proc. Europ. Conf. Speech Commun. Technol.*, Madrid, Spain, Sept. 18–21, 1995, pp. 1559–1562.
- [30] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoustical Society America*, vol. 109, pp. 3007–3020, 2001.
- [31] R. Goecke, G. Potamianos, and C. Neti, "Noisy audio feature enhancement using audio-visual speech data," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Orlando, FL, May 13–17, 2002, pp. 2025–2028.
- [32] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVDCN (audio-visual codebook dependent cepstral normalization)," in *Proc. Int. Conf. Spoken Lang. Processing*, Denver, CO, Sept. 16–20, 2002, pp. 1449–1452.
- [33] C. Bregler and Y. Konig, "'Eigenlips' for robust speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Adelaide, Australia, Apr. 19–22, 1994, pp. 669–672.
- [34] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. Int. Conf. Image Processing*, vol. I, Chicago, IL, Oct. 4–7, 1998, pp. 173–177.
- [35] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, UT, May 7–11, 2001, pp. 177–180.
- [36] J. Luetttin and N. A. Thacker, "Speechreading using probabilistic models," *Computer Vision Image Understanding*, vol. 65, pp. 163–178, 1997.
- [37] D. Chandramohan and P. L. Silsbee, "A multiple deformable template approach for visual speech recognition," in *Proc. Int. Conf. Spoken Lang. Processing*, Philadelphia, PA, Oct. 3–6, 1996, pp. 50–53.
- [38] M. T. Chan, "HMM based audio-visual speech recognition integrating geometric- and appearance-based visual features," in *Proc. Works. Multimedia Signal Processing*, Cannes, France, Oct. 3–5, 2001, pp. 9–14.
- [39] S. Dupont and J. Luetttin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, pp. 141–151, Sept. 2000.
- [40] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–38, Jan. 1998.
- [41] A. W. Senior, "Face and feature finding for a face recognition system," in *Proc. Int. Conf. Audio Video-based Biometric Person Authentication*, Washington, DC, Mar. 22–23, 1999, pp. 154–159.
- [42] J. Huang, G. Potamianos, and C. Neti, "Improving audio-visual speech recognition with an infrared headset," in *ISCA Tut. Res. Works. Audio-Visual Speech Processing*, St. Jorioz, France, Sept. 4–7, 2003.
- [43] P. Teissier, J. Robert-Ribes, and J. L. Schwartz, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 629–642, Nov. 1999.
- [44] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1260–1273, Nov. 2002.
- [45] M. Gordan, C. Kotropoulos, and I. Pitas, "A support vector machine-based dynamic network for visual speech recognition applications," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1248–1259, Nov. 2002.
- [46] A. Rogozan, "Discriminative learning of visual data for audiovisual speech recognition," *Int. J. Artificial Intell. Tools*, vol. 8, pp. 43–52, 1999.
- [47] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Seattle, WA, May 12–15, 1998, pp. 3733–3736.
- [48] 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, Jan. 2000.
- [49] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. Int. Conf. Spoken Lang. Processing*, Philadelphia, PA, Oct. 3–6, 1996, pp. 426–429.
- [50] P. Jourlin, "Word dependent acoustic-labial weights in HMM-based speech recognition," in *Proc. Europ. Tut. Works. Audio-Visual Speech Processing*, Rhodes, Greece, Sept. 27–28, 1997, pp. 69–72.
- [51] S. Nakamura, H. Ito, and K. Shikano, "Stream weight optimization of speech and lip image sequence for audio-visual speech recognition," in *Proc. Int. Conf. Spoken Lang. Processing*, vol. III, Beijing, China, Oct. 16–20, 2000, pp. 20–23.
- [52] S. M. Chu and T. S. Huang, "Audio-visual speech modeling using coupled hidden Markov models," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Orlando, FL, May 13–17, 2002, pp. 2009–2012.
- [53] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1274–1288, Nov. 2002.