

An Image Transform Approach for HMM Based Automatic Lipreading

Gerasimos Potamianos, Hans Peter Graf, and Eric Cosatto

AT&T Labs—Research, 180 Park Ave, Florham Park, NJ 07932-0971, U.S.A.
 email: {makis,hpg,eric}@research.att.com

Abstract

This paper concentrates on the visual front end for hidden Markov model based automatic lipreading. Two approaches for extracting features relevant to lipreading, given image sequences of the speaker's mouth region, are considered: A lip contour based feature approach, which first obtains estimates of the speaker's lip contours and subsequently extracts features from them, and an image transform based approach, which obtains a compressed representation of the image pixel values that contain the speaker's mouth. Various possible features are considered in each approach, and experimental results on a number of visual-only recognition tasks are reported. It is shown that the image transform based approach results in superior lipreading performance. In addition, feature mean subtraction is demonstrated to improve performance in multi-speaker and speaker-independent recognition tasks. Finally, the effects of video degradations to image transform based automatic lipreading are studied. It is shown that lipreading performance dramatically deteriorates below a 10 Hz field rate, and that image transform features are robust to noise and compression artifacts.

1. INTRODUCTION

Automatic recognition of speech by using the video sequence of the speaker's lips, namely *automatic lipreading*, or *speech-reading*, has recently attracted significant interest [1]-[10]. Much of this interest focuses on ways of combining the video channel information with its audio counterpart, in the quest for an *audio-visual* automatic speech recognition (ASR) system that outperforms audio-only ASR [1]-[6]. Such a performance improvement depends on both the audio-visual fusion architecture, as well as on the *visual front end*, namely, on the extraction of appropriate visual features that contain relevant information about the spoken word sequence. In this paper, we concentrate on the latter. We consider a number of visual features, propose new ones, compare them on the basis of lipreading performance, and investigate their robustness to video degradations.

Various visual features have been proposed in the literature that, in general, can be grouped into *lip contour* based and *pixel* based ones [2]. In the first approach, the speaker's lip contours are extracted from the image sequence. A parametric [2], [9] or statistical [10] lip contour model is then obtained, and the model parameters are used as visual features. Alternatively, lip contour geometric features are used [4], [5]. In the second approach, the entire image containing

the speaker's mouth is considered as informative for lipreading, and appropriate transformations of its pixel values are used as visual features [1], [3], [5]-[10].

In this work, we investigate both visual feature approaches. In Section 2, we define a number of lip contour based visual features that our experiments have demonstrated to be successful in lipreading. The use of lip contour Fourier descriptors is novel. In Section 3, we describe our implementation of the pixel based approach. We have considered various linear image transforms for feature extraction, among which the discrete wavelet transform (DWT) [11], the discrete cosine transform [12], and a principal component analysis (PCA) based projection [13] performed the best. The use of the DWT and our implementation of the PCA, based on the correlation matrix of three-dimensional data, are new. In all cases, the extracted visual features are appropriately postprocessed and used in a *hidden Markov model* (HMM) [14] based automatic lipreading system, as described in Section 4. Lipreading performance on the AT&T audio-visual database [5] is reported in Section 5.

Finally, the effects of three types of video degradation on lipreading accuracy are investigated in Section 6, in conjunction with the use of image transform based features. All three, field rate decimation, noise, and compression artifacts, arise in practical video capturing and transmission. To our knowledge, no previous studies on this subject exist, with the exception of [15], where the effects of rate decimation are studied on a simple *human* lipreading task, and [16], where possible effects of video compression on person authentication accuracy are discussed.

2. LIP CONTOUR BASED FEATURES

2.1. Lip Contour Extraction

The lip contour extraction system is described in detail elsewhere [17]. In its current implementation, for each video field, two channels of processing are used: A combination of *shape* and *texture analysis*, and a *color segmentation*, to first locate the mouth and then the precise lip shape. Estimated outer and inner lip contours are depicted in Fig. 1. For a single speaker, part of the outer lip contour is missed in less than 0.25% of the processed images. However, inner lip and multi-speaker contour estimation are less robust.

2.2. Visual Features

Let an outer, or inner, lip contour $\mathcal{C} = \{(x, y)\} \subset \mathcal{R}^2$ be defined as a polygon with "ordered" vertices $\{p_i = (x_i, y_i)$,

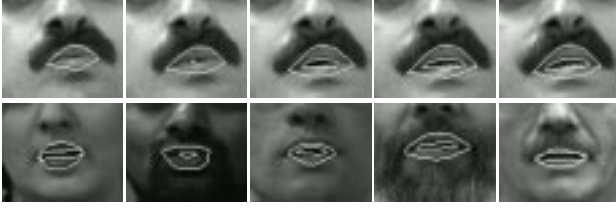


Figure 1: Outer and inner lip contour estimation examples. *Upper row*: Consecutive fields from database part P.1 (see Table 1). *Lower row*: Zoomed in fields of five subjects from the multi-speaker database part P.3.

$i = 1, \dots, n$. Given \mathcal{C} , let us define line l , by

$$y = \frac{\text{Cov}[x, y]}{\text{Var}[x]} x + \text{E}[y] - \frac{\text{Cov}[x, y]}{\text{Var}[x]} \text{E}[x].$$

Line l constitutes the *linear predictor* of the contour y -coordinates, given its x -coordinates. Let also l^\perp be any line perpendicular to l . Then, we define the contour width and height as (see also Fig. 2)

$$w = \max \{ d(\mathcal{P}_l(p_i), \mathcal{P}_l(p_j)) : i, j = 1, \dots, n \}$$

and

$$h = \max \{ d(\mathcal{P}_{l^\perp}(p_i), \mathcal{P}_{l^\perp}(p_j)) : i, j = 1, \dots, n \},$$

respectively, where $\mathcal{P}_l(p)$ denotes the projection of point p on line l , and $d(p, p')$ is the Euclidean distance between points p, p' . In addition, we define the contour perimeter as

$$p = d(p_n, p_1) + \sum_{i=1}^{n-1} d(p_i, p_{i+1}),$$

and the contour area as $a = \mu_{00}$, where [12]

$$\mu_{pq} = \sum_x \sum_y x^p y^q f(x, y), \quad p, q \in \mathcal{N},$$

denote the *moments* of the contour interior binary image

$$f(x, y) = \begin{cases} 1, & \text{if } (x, y) \in \mathcal{C} \cup \mathcal{C}_{\text{interior}}; \\ 0, & \text{otherwise.} \end{cases}$$

Central moments [12], defined as

$$m_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y), \quad p, q \in \mathcal{N},$$

where (\bar{x}, \bar{y}) denotes the contour *center of mass*, with $\bar{x} = \mu_{10}/\mu_{00}$, $\bar{y} = \mu_{01}/\mu_{00}$, as well as, *normalized moments* [12], can also be considered as contour features.

Finally, features obtained from the contour Fourier series representation are considered. Let $(x(t), y(t))$, $t \in [0, T]$, be a parametrization of \mathcal{C} [12], and let the Fourier series expansion of $x(t)$ and $y(t)$ be (see also Fig. 3)

$$x(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos \frac{2n\pi t}{T} + B_n \sin \frac{2n\pi t}{T}$$

and

$$y(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos \frac{2n\pi t}{T} + D_n \sin \frac{2n\pi t}{T},$$

respectively. We compute and subsequently normalize a number of Fourier coefficients (A_n, B_n, C_n, D_n) as described in [12], thus obtaining a set of normalized Fourier coefficients ($A_n^*, B_n^*, C_n^*, D_n^*$). We then consider as contour features the Fourier descriptor magnitudes

$$\mathcal{FD}_n = \sqrt{A_n^{*2} + B_n^{*2} + C_n^{*2} + D_n^{*2}}, \quad \text{with } n \geq 2.$$

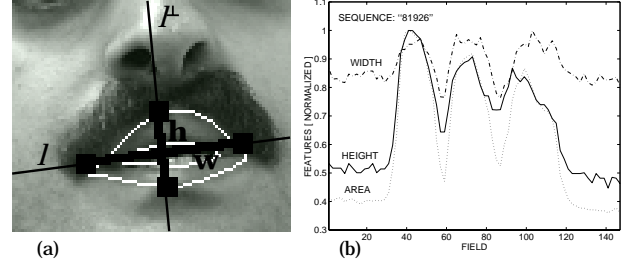


Figure 2: Geometric feature approach. (a): Outer lip width and height. (b): Three geometric visual features, tracked over the spoken sequence "81926", displayed on a normalized scale.

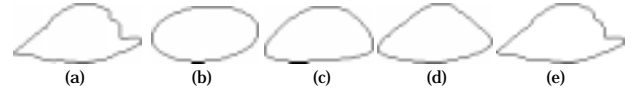


Figure 3: Lip contour Fourier descriptors. (a): Estimated outer lip contour \mathcal{C} . (b)-(e): Reconstructed \mathcal{C} from 1, 2, 3, and 20 sets of Fourier coefficients, respectively.

3. IMAGE TRANSFORM BASED FEATURES

3.1. Image Sequence Preprocessing

In the pixel based approach to feature extraction, a *region of interest* (ROI) containing the speaker's mouth is first defined. In our system, each YUV422 video field, available at a 60 Hz rate, is processed to obtain the outer lip contour estimate \mathcal{C} . The Y-band image of the field is then retained, and *optionally* normalized with an *affine* transformation (based on \mathcal{C}) and histogram equalization. Such normalization can somewhat compensate for head rotation, camera-subject distance, and lighting variations. Finally, a 64×64 pixel window around the contour center of mass (\bar{x}, \bar{y}) is obtained, and further downsampled to 16×16 pixels. The resulting video sequence is denoted by $\{g(x, y, t) : 1 \leq x \leq M, 1 \leq y \leq N, 1 \leq t \leq T\}$, where $M = N = 16$, and T is the number of video sequence fields (see Fig. 4).

3.2. Image Sequence Transforms

Given the preprocessed video sequence, at every time instance t we consider as data relevant to lipreading the MNK pixels in $\{g(x, y, z) : 1 \leq x \leq M, 1 \leq y \leq N, t - t_L \leq z \leq t + t_R\}$, where $t_L, t_R \geq 0$, and $K = t_L + t_R + 1$. Let us lexicographically order this data into a MNK -element data vector $\underline{g}_t = [g_{t,1}, \dots, g_{t,MNK}]'$. We seek a *linear transform* matrix $\underline{\mathbf{P}} = [\underline{P}_1, \dots, \underline{P}_{MNK}]$, such that $\underline{g}'_t \underline{\mathbf{P}}$ contains information relevant to lipreading in only few of its elements. An operator \mathcal{S} then places the J most informative such elements into a J -element feature vector $\underline{o}_t = \mathcal{S}[\underline{g}'_t \underline{\mathbf{P}}]$. In order to obtain matrix $\underline{\mathbf{P}}$ and operator \mathcal{S} , L training examples are given, that we denote by \underline{g}_l , $l = 1, \dots, L$.

3.2.1. Discrete Wavelet, Cosine, and Walsh Transforms

A number of linear, *separable* image transforms can be used in place of $\underline{\mathbf{P}}$. In this work, we consider the *discrete wavelet* (DWT) [11], the *discrete cosine* (DCT), and the *Walsh* (WAL) transforms [12]. The DWT is implemented by means of the Daubechies class wavelet filter of approximating order 3 with filter coefficients $[0.333, 0.807, 0.460,$

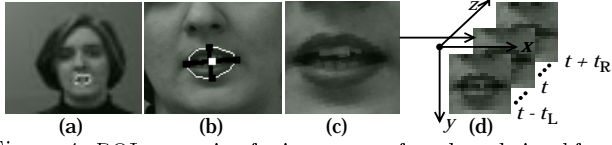


Figure 4: ROI extraction for image transform based visual front end. (a): Original video field with estimated outer lip contour and center of mouth. (b): Speaker's mouth region, zoomed in. (c): Video field ROI after an optional normalization step. (d): Final, downsampled, three-dimensional ROI at time t .

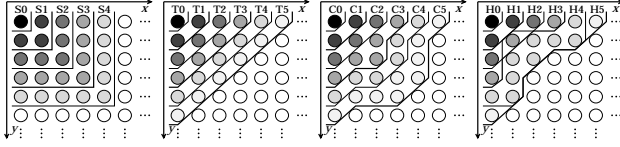


Figure 5: Various planar sublattices Λ , where image transform coefficients are retained. "S", "T", "C", and "H" denote square, triangular, circular, and hyperbolic sublattices, respectively.

$-0.135, -0.085, 0.035$] [11], [18]. Fast algorithms for all three transforms exist and are used in our case, since we consider $M=N=16$, and $K=1, 2, 4, 8$.

To design operator \mathcal{S} , given the set of training examples, we compute the average energy of the transformed vector on each of the MNK sites as

$$E_k = \frac{1}{L} \sum_{l=1}^L \langle \underline{g}_l, \underline{P}_k \rangle^2, \text{ for } k = 1, \dots, MNK,$$

where $\langle \bullet, \bullet \rangle$ denotes vector *inner product*. Let the $J \ll MNK$ largest energies be $\{E_{k_1}, \dots, E_{k_j}\}$. Then, given a data vector \underline{g}_t , we obtain its feature vector as

$$\underline{o}_t = [o_{t,1}, \dots, o_{t,J}]', \text{ where } o_{t,j} = \langle \underline{g}_t, \underline{P}_{k_j} \rangle.$$

Alternatively, the transformed vector elements that lie on an a-priori, appropriately chosen sublattice Λ are retained. Various such sublattices are depicted in Fig. 5, for $K=1$.

3.2.2. Principal Component Analysis

Principal component analysis (PCA) [12], [13] achieves the optimal information compression, in the sense of minimum mean square error between \underline{g}_t and its reconstruction based on \underline{o}_t . PCA projects the data onto the directions of their greatest variance. However, the problem of *scaling* [13] arises in practical applications. In our experiments, we have first considered the PCA projection based on the data *covariance* matrix, giving rise to the *Karhunen-Loève* transform (KLT). However, we found it beneficial to scale our data according to their inverse variance. Namely, we compute the data mean and variance as

$$m_k = \frac{1}{L} \sum_{l=1}^L g_{l,k}, \text{ and } \sigma_k^2 = \frac{1}{L} \sum_{l=1}^L (g_{l,k} - m_k)^2,$$

for $k = 1, \dots, MNK$, respectively, and the data *correlation* $MNK \times MNK$ matrix \mathbf{R} with elements

$$r_{k,k'} = \frac{1}{L} \sum_{l=1}^L \frac{(g_{l,k} - m_k)}{\sigma_k} \frac{(g_{l,k'} - m_{k'})}{\sigma_{k'}},$$

for $k, k' = 1, \dots, MNK$. We then *diagonalize* the correlation matrix as $\mathbf{R} = \mathbf{A}\mathbf{D}\mathbf{A}'$ [13], [18], where $\mathbf{A} = [\underline{a}_1, \underline{a}_2, \dots, \underline{a}_{MNK}]$ has as columns the *eigenvectors* of \mathbf{R} , and \mathbf{D} is a diagonal

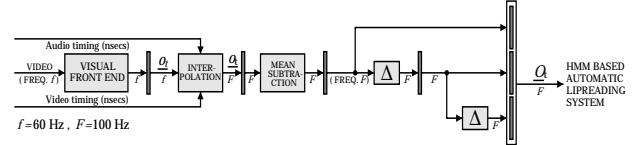


Figure 6: Visual feature postprocessing (see Section 4.1).

matrix containing the *eigenvalues* of \mathbf{R} . Let the J largest such eigenvalues be $\{\lambda_{k_1}, \dots, \lambda_{k_j}\}$, where $\lambda_{k_j} = \mathbf{D}_{k_j k_j}$. Given a data vector \underline{g}_t , we first normalize it into

$$\underline{G}_t = [G_{t,1}, \dots, G_{t,MNK}]', \text{ where } G_{t,k} = \frac{g_{t,k} - m_k}{\sigma_k},$$

and then extract its J -element feature vector as

$$\underline{o}_t = [o_{t,1}, \dots, o_{t,J}]', \text{ where } o_{t,j} = \langle \underline{G}_t, \underline{a}_{k_j} \rangle.$$

4. THE HMM BASED LIPREADING SYSTEM

4.1. Visual Feature Postprocessing

For every video field, a "static" observation feature vector \underline{o}_t is acquired, as described in Sections 2, or 3. A number of *linear* operations is then applied to \underline{o}_t . First, in an audio-visual system, the sequence of \underline{o}_t , $1 \leq t \leq T$, can be aligned to the sequence of audio observation vectors that are typically produced every 10 ms [14], by means of *linear interpolation* [18]. This facilitates visual-only HMM training (as well as, "early integration" based audio-visual speech recognition [2], [6]). The resulting static visual feature vector is \underline{o}_t . In addition, the feature mean over the spoken sequence is *subtracted*, to improve multi-speaker and speaker-independent lipreading performance. Finally, discrete approximations of first (Δ) and second *derivatives* of \underline{o}_t , over time, are computed. The final observation vector used in the HMM based lipreading system is $\underline{O}_t = [\underline{o}_t - \mathbf{E}[\underline{o}_t], \Delta \otimes \underline{o}_t, \Delta \otimes \Delta \otimes \underline{o}_t]$ (see Fig. 6).

4.2. HMM Specifics

Our automatic lipreading system uses continuous density HMMs as a means of statistical pattern matching. The HMM observation probabilities are modeled as multi-dimensional Gaussian mixtures with diagonal covariance matrices [14]. For the specific lipreading recognition tasks considered in this paper (see Table 1), we use context independent, whole word, 6-10 state, left-to-right models with 16 mixtures per state, and a single state silence model with 32 mixtures. All HMM parameters are estimated by maximum likelihood Viterbi training [14], with the exception of some results in Table 4, where a discriminative training algorithm is used (generalized probabilistic descent [14]). The initial segmentation in the training procedure is obtained by using the audio channel information and available audio-only HMMs with identical structure to their visual counterparts. *Unknown string length* is assumed at recognition.

5. AUTOMATIC LIPREADING RESULTS

The three database parts and the testing scenarios are depicted in Tables 1 and 2, respectively. We first consider the use of geometric features on parts P.1 and P.2. Outer lip

Part	Subjects	Task	Voc.	Words
P.1	1	digit strings	11	600 × 5
P.2	1	letter strings	26	2500 × 4
P.3	49	letter strings	26	1225 × 4

Table 1: Three parts of the AT&T bimodal database considered in this paper. Number of subjects, task, vocabulary size, and number of words (sentences × words per sentence) are shown.

Part	Scenario	Training set	Test set
P.1	I.S.	1 × 400 × 5	1 × 200 × 5
P.2	I.S.	1 × 2000 × 4	1 × 500 × 4
P.3	M.S.	49 × 20 × 4	49 × 5 × 4
P.3	S.I.	40 × 25 × 4	9 × 25 × 4

Table 2: Testing scenarios for the three database parts. A multi-speaker (M.S.) and a speaker-independent (S.I.) scenario are considered in part P.3, whereas both parts, P.1 and P.2, are single-speaker (I.S.). Training and test set sizes are shown as number of speakers × sentences per speaker × words per sentence.

contours are significantly more reliable than inner lip ones. Therefore, their features result in better lipreading than those based on inner lip contours (see Table 3). In addition, incorporating their Fourier descriptors into the feature vector results in improved recognition. Using both outer and inner lip contour features improves lipreading. No results are reported on part P.3, where lip contour estimation is less reliable.

We then consider image transform based visual features. Performance of a number of DWT and PCA based features on part P.1 is depicted in Fig. 7. PCA features provide a more compact representation (smaller J), whereas, the DWT typically requires higher-dimensional feature vectors to reach optimal performance. In Table 4, DWT, DCT, WAL, KLT, and PCA based lipreading results on all database tasks are depicted. Our experiments indicate that small values of $K > 1$ are typically beneficial to WAL, KLT, and PCA based approaches (this is depicted in the PCA case). In addition, the KLT (i.e., PCA, based on data covariance) performs consistently worse than the PCA based on data correlation. Although the relative performance of the various methods differs across tasks, the DWT (as well as, the DCT) based features with $K=1$ and $J=17$ perform consistently well. In the last line of Table 4, their lipreading performance, achieved by means of discriminatively trained HMMs, is depicted.

Clearly, our results cannot be strictly compared with lipreading performance reported elsewhere in the literature, on different databases. Nevertheless, they measure well against the 77% word accuracy reported in [7] on a P.1-type task, as well as, the 53% word accuracy reported in [1] on a German P.2-type task. In the revised version of [3], a 34.2% word accuracy is reported on a multi-speaker *isolated* letter task (a simpler task than P.3-M.S.).

Some additional experimental results are worth reporting: Feature mean subtraction significantly helped recognition in both P.3 tasks. Indeed, using the suggested DWT features, it improved word accuracy from 10.5% to 29.6% and from 4.4% to 19.6% on tasks P.3-M.S. and P.3-S.I., respectively, although it did not significantly affect performance on single-speaker tasks P.1 and P.2. Surprisingly,

Part	Outer lip \mathcal{C}_O	Inner lip \mathcal{C}_I	Both $\mathcal{C}_O, \mathcal{C}_I$
P.1	73.4% (19.5%)	64.0% (13.5%)	83.9% (44.0%)
P.2	31.2% (1.0%)	24.2% (0.2%)	42.9% (4.8%)

Table 3: Lipreading performance in word (string) accuracy on database parts P.1 and P.2, based on lip contour features w, h, a, p (both $\mathcal{C}_O, \mathcal{C}_I$) and $\mathcal{F}\mathcal{D}_2\text{-}\mathcal{F}\mathcal{D}_5$ (\mathcal{C}_O only).

P, K, J/Λ	P.1	P.2	P.3-M.S.	P.3-S.I.
DWT,1,H4	92.4(66.5)	59.2(12.8)	26.3(0.0)	18.6(0.9)
DWT,1,17	92.4(67.5)	57.4(11.8)	29.6(1.6)	19.6(0.4)
DCT,1,H4	93.2(70.5)	57.2(11.2)	29.2(1.2)	19.7(0.4)
DCT,4,17	91.8(66.0)	55.8 (9.8)	28.6(0.8)	19.6(0.4)
WAL,4,17	91.3(63.0)	52.2 (7.6)	27.4(0.4)	17.0(0.9)
KLT,1,10	88.0(51.5)	51.0 (7.2)	24.5(0.8)	16.9(0.0)
PCA,1,10	91.6(65.5)	51.4 (7.4)	28.8(1.2)	19.2(0.4)
PCA,3,10	92.2(70.0)	52.9 (9.0)	31.3(0.8)	18.9(0.0)
DWT,1,17	95.7(80.0)	64.5(21.2)	30.3(2.9)	21.0(0.9)

Table 4: Lipreading performance in word (string) % accuracy on all database parts, by means of various image transforms. Last line depicts results obtained by HMM discriminative training.

and unlike in [8], [9], image normalization and use of color information in PCA fail to consistently improve our system’s lipreading performance. Finally, using the combination of both contour based and image transform (DWT) based features degrades lipreading performance on tasks P.1, P.2, as compared to the use of the latter only features.

6. VIDEO QUALITY AND LIPREADING

Often, practical video capturing and transmission requirements result in low field rate, compressed, and noisy video. Therefore, it is of interest to study the effect of these degradations to image transform based automatic lipreading.

(a) *Video field rate decimation*: The field rate of the database is 60 Hz, therefore, we can easily consider lower rates. In Fig. 8a, it becomes clear that use of a 60 Hz video field rate versus a 30 Hz one has a marginal benefit only. The limit for “acceptable” automatic lipreading performance seems to be in the range of 10-15 Hz.

(b) *Video noise*: We contaminate the video fields with

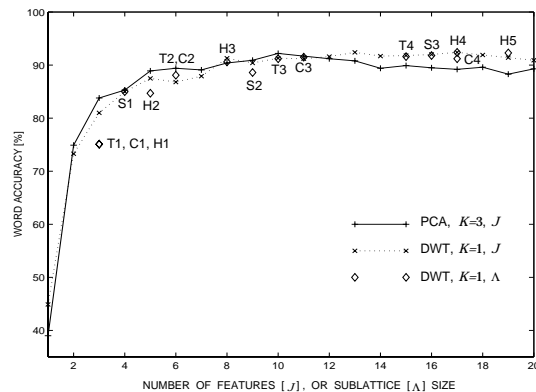


Figure 7: Lipreading performance on part P.1, for various planar sublattices Λ of Fig. 5 using DWT ($K=1$), and for various number J of DWT ($K=1$) and PCA ($K=3$) based features.

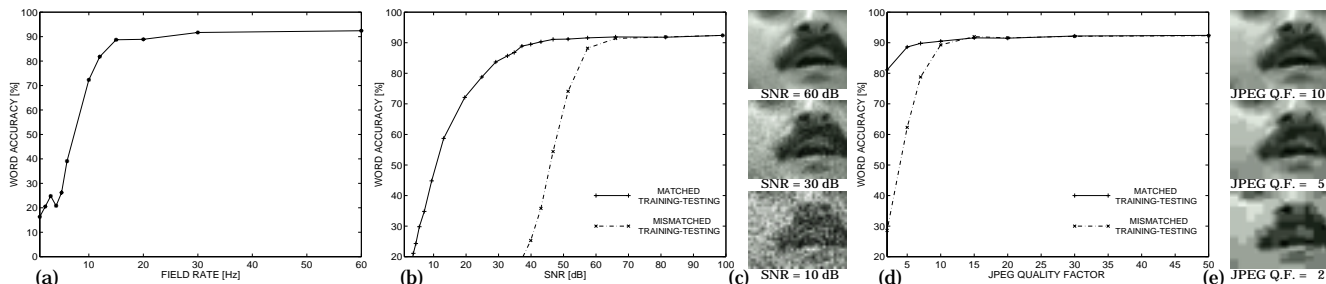


Figure 8: Video degradation effects on lipreading performance on part P.1, by means of DWT features ($K=1$, $\Lambda=H4$). (a): Effect of video field rate. (b): Effect of additive white noise of various SNR values, under matching and mismatched (HMMs are trained on clean video) training and testing. (c): Degraded video fields at various SNRs. (d): Effect of JPEG compression on video fields as a function of JPEG quality factor (Q.F.). (e): Degraded video fields for various JPEG quality factors.

additive white noise at various SNRs, assuming that the center of mouth estimation is not affected. Under matching training and testing conditions, DWT based visual features are quite robust (see Fig. 8b,c).

(c) *Video compression*: Popular still image and video coders, such as JPEG [19] and MPEG2, exhibit mainly blocking artifacts at low bitrates. In Fig. 8d, we consider the effects of JPEG coding applied on still fields. Lipreading performance remains practically unaffected.

7. SUMMARY AND DISCUSSION

In this paper, both lip contour and image transform based visual features are considered for HMM based automatic lipreading. The latter are shown to perform significantly better, while being robust to video degradations, and they result in high visual-only recognition performance on single-speaker, multi-speaker, and speaker-independent tasks.

The superiority of image transform based features is not surprising, since significant speechreading information lies within the oral cavity that cannot be captured by the lip contours. In addition, lip contour estimation errors compromise the recognition accuracy. As is clear from our experiments, a number of image transform based features result in similar lipreading performance. Given the fact that PCA requires an intensive training phase and is not amenable to fast implementation, the use of DWT or DCT based features is recommended. Finally, the demonstrated robustness of image transform features to video degradations is promising for low-cost lipreading system implementation.

References

- [1] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading", *Proc. Int. Conf. Speech Lang. Process.*, Yokohama, 1994.
- [2] 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.
- [3] 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.
- [4] P. Jourlin, "Word-dependent acoustic-labial weights in HMM-based speech recognition", *Proc. Europ. Tut. Work. Audio-Visual Speech Process.*, Rhodes, pp. 69-72, 1997.
- [5] G. Potamianos, E. Cosatto, H.P. Graf, and D.B. Roe, "Speaker independent audio-visual database for bimodal ASR," *Proc. Europ. Tut. Work. Audio-Visual Speech Proc.*, Rhodes, pp. 65-68, 1997.
- [6] 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.
- [7] N.M. Brooke, "Talking heads and speech recognizers that can see: The computer processing of visual speech signals", in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 351-371, 1996.
- [8] M.S. Gray, J.R. Movellan, and T.J. Sejnowski, "Dynamic features for visual speech-reading: A systematic comparison", in *Advances in Neural Information Processing Systems 9*, M.C. Mozer, M.I. Jordan, and T. Petsche eds., MIT Press, Cambridge, pp. 751-757, 1997.
- [9] G.I. Chiou and J.-N. Hwang, "Lipreading from color video", *IEEE Trans. Image Process.*, vol. 6, pp. 1192-1195, 1997.
- [10] J. Luettin, "Towards speaker independent continuous speechreading," *Proc. Eurospeech*, Rhodes, pp. 1991-1994, 1997.
- [11] I. Daubechies, *Wavelets*. S.I.A.M., Philadelphia, 1992.
- [12] E.R. Dougherty and C.R. Giardina, *Image Processing - Continuous to Discrete, Vol. 1. Geometric, Transform, and Statistical Methods*. Prentice Hall, Englewood Cliffs, 1987.
- [13] C. Chatfield and A.J. Collins, *Introduction to Multivariate Analysis*. Chapman and Hall, London, 1980.
- [14] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, 1993.
- [15] J.J. Williams, J.C. Rutledge, D.C. Garstecki, and A.K. Katsaggelos, "Frame rate and viseme analysis for multimedia applications," *Proc. IEEE Works. Multimedia Signal Process.*, Princeton, pp. 13-18, 1997.
- [16] F. Davoine, H. Li, and R. Forchheimer, "Video compression and person authentication," in *Audio- and Video-based Biometric Person Authentication*, J. Bigün, G. Chollet, and G. Borgefors, eds., Springer, Berlin, pp. 353-360, 1997.
- [17] H.P. Graf, E. Cosatto, and G. Potamianos, "Robust recognition of faces and facial features with a multi-modal system," *Proc. Int. Conf. Systems Man Cybern.*, Orlando, pp. 2034-2039, 1997.
- [18] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes in C. The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1988.
- [19] G.K. Wallace, "The JPEG still picture compression standard", *Commun. ACM*, vol. 34, no. 4, pp. 30-44, 1991.