

IMPROVED ROI AND WITHIN FRAME DISCRIMINANT FEATURES FOR LIPREADING

Gerasimos Potamianos and Chalapathy Neti

IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

E-mails: {gpotam, cneti}@us.ibm.com

ABSTRACT

We study three aspects of designing appearance based visual features for automatic lipreading: (a) The choice of the video region of interest (ROI), on which image transform features are obtained; (b) The extraction of speech discriminant features at each frame; and (c) The use of temporal information to improve visual speech modeling. In particular, with respect to (a), we propose a ROI that includes the speaker's jaw and cheeks, in addition to the traditionally used mouth/lip region; with respect to (b) and (c), we propose the use of a two-stage linear discriminant analysis, both within frame, as well as across a large number of frames. On a large-vocabulary, continuous speech audio-visual database, the proposed visual features result in a 13% absolute reduction in visual-only word error rate over a baseline visual front end, and in an additional 28% relative improvement in audio-visual over audio-only phonetic classification accuracy.

1. INTRODUCTION

The use of visual, mouth region information has been considered by many researchers as a means to improve *automatic speech recognition* (ASR) robustness, and thus lead to more natural human-computer interaction [1]-[11]. Such work has been motivated by human perception studies and by the well-known ability of humans to *lipread* (*speechread*) [12]-[16]. Not surprisingly, information extracted from the speaker's video has been shown to improve ASR in both clean and noisy audio conditions, and for a number of recognition tasks [2], ranging from single-speaker, small-vocabulary, isolated-word tasks [3], [4], to speaker-independent small-vocabulary connected-word tasks [5]-[7], and, very recently, to *speaker-independent* (SI), *large-vocabulary*, *continuous speech recognition* (LVCSR) [8]. There exist two key issues for the successful design of audio-visual ASR systems [2]: First, the choice of appropriate *visual features* (*visual front end*), and second, the design of the audio-visual information *fusion* algorithm. In this paper, we exclusively address the first problem.

Various sets of visual features for *automatic speechreading* have been proposed in the literature over the last 20 years. In general, they can be grouped into three categories: High-level *lip contour* (*shape*) based features, low-level *video pixel* (*appearance*) based ones, and features that are a combination of both [2]. In the first approach, the speaker's inner and (or) outer lip contours are extracted from the image sequence. A parametric, or statistical lip contour model is then obtained, and the model parameters are used as visual features [2]. Alternatively, lip contour geometric features are used, such as mouth height and width [3], [4]. In the second approach, the entire image containing the speaker's mouth is considered as informative for speechreading (*region of interest* - ROI), and appropriate *image transformations* of its pixel values are used as visual features [5]-[11]. Often, the high- and low-level feature extraction approaches are combined to give rise to joint shape and appearance visual features [7], [8], [10].

Among the above approaches, the appearance (image transform) based features are the most efficient, since they typically

employ a "gross" face-tracking system to extract the approximate ROI containing the speaker's mouth. In contrast, the other two approaches require "expensive" lip and, possibly, face contour modeling and tracking. Furthermore, image transform based features have been shown to outperform lip-contour and shape-model based ones in [8], [10], [11]. A simple and computationally efficient image transform is the two-dimensional, separable, *discrete cosine transform* (DCT) [5]. DCT based visual features for automatic speechreading have been experimentally shown to perform equally well or better than alternative image transform features, such as discrete wavelet transform (DWT), or principal component analysis (PCA) based ones [9]-[11]. Therefore, the DCT has been used as the basis of the visual front end in our previous work [8], [9]. There, and in order to improve visual speech modeling, dynamic (temporal) information is also incorporated into feature extraction, by considering the concatenation of a number of neighboring DCT features, followed by a *linear discriminant analysis* (LDA) based data projection [17], and a *maximum likelihood linear transform* (MLLT) that amounts to a feature vector rotation [18]. The resulting visual features provide sufficient speech information to improve LVCSR in both clean and noisy audio conditions [8].

However, a number of issues have not been investigated in this baseline visual front end design [9]. In particular, of great interest is the question of *what part of the face should the visual ROI include*. In [9], the ROI has been chosen as a square, centered at the estimated mouth center, and wide enough to include the size-normalized speaker's mouth, thus incorporating both the lip-region and the oral cavity. However, a number of human audio-visual speech perception studies have indicated that an augmented ROI, containing the entire lower half of the speaker's face (including the cheeks and the jaw), is beneficial to human lipreaders [13], [14]. To our knowledge, there exists no similar study on the effects of ROI selection to automatic speechreading. Other human perception studies have demonstrated that temporal visual information that spans multiple phone segments is very useful in human lipreading [16]. This fact has also not been sufficiently explored in visual feature extraction for an automatic speechreading system.

In this paper, motivated by the above studies, we investigate the effects to automatic speechreading of both ROI selection, as well as of the temporal window size, when extracting visual features. In particular, we demonstrate that using the entire lower half of the subject's face as the ROI improves speechreading significantly, over using the mouth-only ROI. Of course, the augmented ROI presents the additional challenge of selecting speech informative features within each video frame. We improve the DCT energy based feature selection of our baseline system by introducing a new LDA/MLLT data projection/rotation for extracting discriminative, frame-level features. We demonstrate that this step improves speechreading performance. Finally, we show that long temporal windows also improve performance at a significant however increase in computation. Similarly to the baseline system, LDA and MLLT are also applied to the concatenation of neighboring frame-level features, giving rise to the final visual feature vector. The entire visual front end thus amounts to a two stage

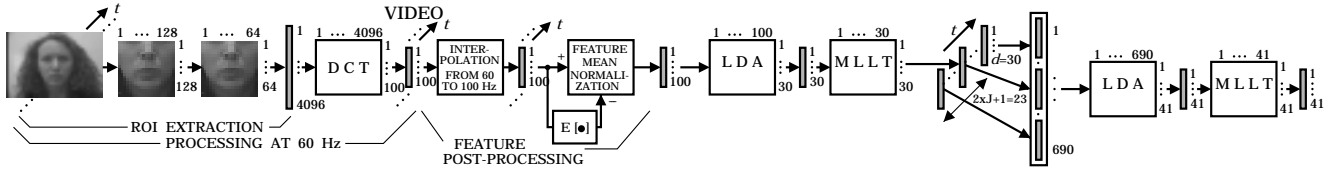


Fig. 1. The proposed visual front end for automatic speechreading. It augments our baseline system of [9], by using within frame LDA/MLLT feature extraction, by considering a larger ROI that contains the lower half face, and by applying a second LDA/MLLT on a longer temporal window.

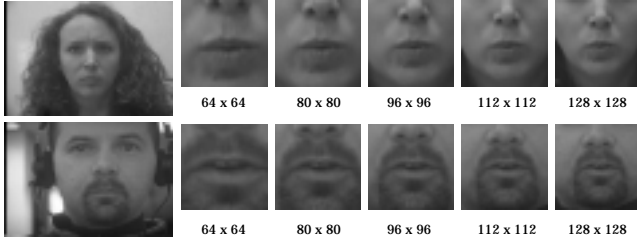


Fig. 2. Original video frames, baseline system extracted ROIs ($m = n = 64$), and the augmented ROIs ($m = n = 80, 96, 112, 128$), considered here. *Upper row*: LVCSR data; *Lower row*: DIGITS data (see Table 1).

application of LDA and MLLT on high energy DCT coefficients, first within frames, and subsequently, across frames.

This paper is structured as follows: In Section 2, we discuss the baseline visual front end system of [9]. In Section 3, we propose three improvements to it: Lower half face ROI extraction; discriminant, within frame visual features; and a longer temporal window for modeling visual speech dynamics. Our experimental results are presented in Section 4, and our conclusions are drawn in Section 5.

2. THE BASELINE VISUAL FRONT END

The baseline DCT based visual front end is described in detail in [8], [9] (see also Fig. 1). Briefly, a statistical face tracking algorithm is first used to detect the speaker's face and to estimate the mouth location and size [9]. Based on these, a $m \times n$ -pixel (where $m = n = 64$), size-normalized ROI is extracted for every video frame at 60 Hz, that contains the speaker's mouth (see also Fig. 2). Subsequently, a two-dimensional, separable, DCT is applied to the ROI, and the $D = 24$ highest energy (over all training data) DCT coefficients are retained as static, frame-level features. To facilitate audio-visual fusion, linear interpolation is used to obtain visual features, time-synchronous to the audio ones at 100 Hz. Utterance-level, *feature mean normalization* (FMN) is employed to compensate for lighting variations, providing the final static features of dimension $d = D = 24$. A number of $J = 7$ neighboring static such features at each side of the current frame are then concatenated to it, giving rise to a dynamic feature vector of dimension $(2 \times J + 1) \times d$. Subsequently, LDA is applied to this vector, reducing its dimensionality in a discriminative manner (given the classes of interest; in our case, the context dependent hidden Markov model states of an ASR system are used as classes; see also Section 4). Finally, an MLLT data rotation improves maximum likelihood data modeling under the assumption of the data class-conditional Gaussian distribution with diagonal covariance [18]. The resulting visual feature vector is of dimension 41.

3. THE IMPROVED VISUAL FRONT END

As mentioned in the introduction, the proposed visual front end differs in three aspects from the baseline system. A more detailed description follows.

Task	Training (U/S/D)	Adaptation	Testing
LVCSR-Ph	5000/239/10.4	N / A	500/26/1.1
LVCSR-SI	17111/239/34.9	N / A	1038/26/2.5
LVCSR-SA	LVCSR-SI(train)	855/26/2.1	LVCSR-SI(test)
DIGITS-SA	LVCSR-SI(train)	120/3/0.2	20/3/0.02

Table 1. Acronyms for the recognition tasks considered in this paper. Number of utterances (U), number of subjects (S), and data duration in hours (D) of the training, test, and adaptation (if applicable) sets are depicted. *LVCSR-Ph*: Speaker-independent visual-only and audio-visual phonetic classification on a subset of the VVAV data. *LVCSR-SI*: Speaker-independent visual-only ASR (LVCSR) for the entire VVAV database. *LVCSR-SA*: Speaker-adapted (per-subject, MLLR) visual-only ASR on the same database. *DIGITS-SA*: Connected digits visual-only recognition on a small dataset collected in a mismatched environment to the VVAV dataset, after per-speaker MLLR adaptation of HMMs, trained on the VVAV data.

3.1. Region of Interest Extraction

Similarly to the baseline system, we center the square, $m \times n$ -pixel ROI at the speaker's mouth center, and we normalize the frame based on the mouth size. However, instead of extracting a ROI with $m = n = 64$ (which includes the mouth region only), we consider a number of successively larger ROIs, with $m = n = 80, 96, 112$, and 128 . Example ROIs are depicted in Fig. 2. Clearly, the largest two ROIs include most of the speaker's jaw and cheeks, which contain useful speech information [13], [14]. All ROIs are subsampled back to a 64×64 pixel size, keeping the dimensionality of all ROIs constant to 4096 pixels. Such subsampling is not expected to seriously affect speechreading performance (see also [15]). Based on the experiments reported in Section 4, using the largest two ROIs ($m = n = 112$, or 128), improves visual speech recognition significantly. Thus, the proposed visual front end uses $m = n = 128$ (performance, when $m = n = 112$, is similar).

3.2. Discriminant Frame-Level Feature Selection

In the baseline system, the first few highest energy DCT coefficients of the ROI are considered as the most informative about visual speech. However, this need not be true, especially in the case of the augmented ROIs. In this work, we propose to keep a larger number of DCT coefficients, D , compared to the baseline system, and to use a data driven approach to obtain discriminant, frame-level features of a lower dimension, d . Such dimensionality reduction is achieved by an additional LDA step, introduced immediately after feature mean normalization (see also Fig. 1). LDA is subsequently followed by an MLLT data rotation. Based on our experiments, values $D = 100$ and $d = 30$ have been selected for the improved visual front end (values $D = 50$ and $d = 30$ result in similar speechreading performance).

3.3. Temporal Window Size

Similarly to the baseline system, we use a concatenation of neighboring frame-level features to obtain dynamic, visual speech in-

Front End Parameters				Modality	
$m \times n$	D	d	J	VI	AV
64×64	24	24	7	26.49	55.19
64×64	50	24	7	27.10	55.40
64×64	50	30	7	27.26	55.37
64×64	100	30	7	27.64	55.60
64×64	50	30	7	27.26	55.37
64×64	50	30	8	27.71	55.74
64×64	50	30	9	27.83	55.57
64×64	50	30	10	27.98	55.64
64×64	50	30	11	28.34	55.79
64×64	50	30	12	28.49	55.88
64×64	50	30	13	28.62	55.94
64×64	100	30	7	27.64	55.60
80×80	100	30	7	28.15	55.79
96×96	100	30	7	28.60	55.85
112×112	100	30	7	29.19	56.01
128×128	100	30	7	29.25	56.03
64×64	24	24	7	26.49	55.19
128×128	100	30	11	30.36	56.45

Table 2. Visual-only (VI) and audio-visual (AV) phonetic classification accuracy (%) on the LVCSR-Ph task (see Table 1), using various visual front end parameters; namely, a varying ROI size (m, n), number of highest energy DCT coefficients at the frame level (D), dimensionality of frame-level features after the first stage of LDA/MLLT (d), and temporal window size ($2 \times J + 1$). Final visual features are always of dimension 41. Comparisons are grouped in blocks. The baseline and the proposed visual front end performances are given in the last two lines. Audio-only accuracy is 50.63% (see also Fig. 3).

formation. A second stage of LDA/MLLT is subsequently used to derive discriminant visual features of the same reduced dimensionality as the baseline system, namely 41. Our experiments indicate that longer temporal windows consistently improve performance, in agreement with human perception studies [16]. However, computing the LDA matrix, a task of $O(J^3)$ complexity, becomes prohibitive for high input vector dimensions, while at decoding, the LDA matrix multiplication is a computationally expensive task of $O(J^2)$ cost. We, therefore, use value $J = 11$ in the proposed visual front end. This corresponds to concatenating 23 consecutive static frames of a total dimension of $d \times (2 \times J + 1) = 690$. Of course, a real-time visual front end requires lower values of J , or just augmenting the static vector by its first- and possibly second-order temporal derivatives, instead of applying the LDA projection.

4. DATABASE AND EXPERIMENTAL RESULTS

To study the effects of the proposed visual front end to automatic speechreading, we conduct experiments on two datasets. The first is the IBM ViaVoiceTM audio-visual (VVAV) database, described in detail in [8] (39.5 hrs, 265 subjects, large-vocabulary read speech - LVCSR task), and the second is a small dataset of 3 subjects wearing a head-mounted microphone, recently collected in mismatched conditions to the VVAV set (see also Fig. 2). The latter set contains 97 read, continuous speech utterances and 43 connected digit strings.

We first report visual-only and audio-visual *phonetic classification* experiments on a subset of the VVAV database (LVCSR-Ph task), followed by speaker-independent (LVCSR-SI task) and

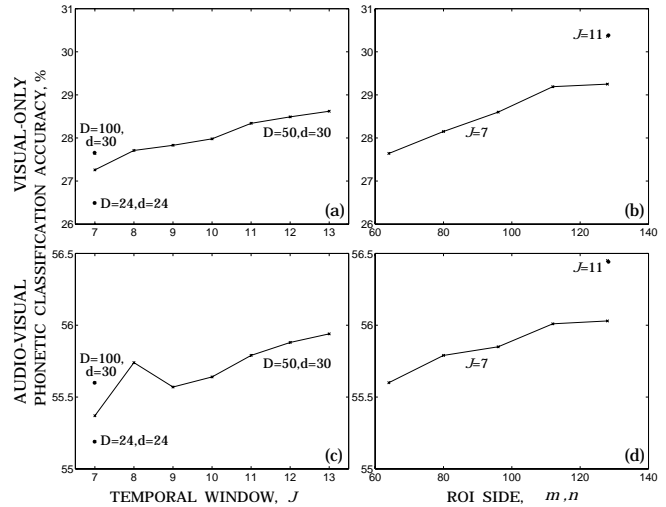


Fig. 3. Visual-only ((a), (b)) and audio-visual ((c), (d)) phonetic classification accuracy (%) on the LVCSR-Ph task (see Table 1), as a function of the following visual front end parameters: *Left* ((a), (c)): Temporal window size ($2 \times J + 1$), when $m = n = 64$; *Right* ((b), (d)): ROI side ($m = n$), when $D = 100, d = 30$. The results are also listed in Table 2.

speaker-adapted (LVCSR-AD task) visual-only ASR experiments on the entire VVAV set (see also Table 1). On the smaller, mismatched dataset, we report speaker-adapted recognition on a test set, consisting of only 20 connected digit strings (DIGITS-SA task), while using the remaining 23 digit and all 97 continuous speech utterances for speaker adaptation of *hidden Markov models* (HMMs) trained on the VVAV database (see Table 1). All adaptation experiments are performed using per-subject HMM adaptation, by means of *maximum likelihood linear regression* (MLLR) [19].

In Table 2, we report visual-only and audio-visual frame-based phonetic classification performance (in accuracy, %) on a VVAV subset (LVCSR-Ph task), using various visual-front end parameters. Similarly to [9], we consider 52 *phoneme* classes with a *uniform* class prior, and, for single-modality classification, we use a *Gaussian mixture model* (GMM) classifier with 5-10 mixtures per class, that model the class-conditional feature observation densities. For audio-visual (bimodal) classification, we first concatenate the visual features to their time-synchronous (due to the interpolation module of Section 2) audio *mel-frequency cepstral coefficient* (MFCC) based features (see [9]), and subsequently, we employ the *multi-stream* (here, two-stream) GMM classifier on the bimodal features [9]. As Table 2 demonstrates, all three visual front end improvements described in Sections 3.1, 3.2, and 3.3, and studied in the third, first, and second block of Table 2, respectively, introduce significant gains in both visual-only and audio-visual performance. The proposed new visual front end improves the baseline system visual-only phonetic classification from 26.49% to 30.36% (a relative 15% gain), and the absolute audio-visual performance gain over the 50.63% audio-only accuracy from 4.56% to 5.82% (a relative 28% gain). These results are also depicted in Fig. 3.

In Table 3, we report visual-only ASR, in *word error rate* (WER) for the speaker-independent and speaker-adapted LVCSR tasks, as well as speaker-adapted performance on the small connected digits (DIGITS-SA) task, for a number of visual front end parameters. All results are obtained using an HMM classifier that contains 2.8 K context-dependent sub-phonetic classes (states), a total of about 47.2 K Gaussian mixtures, and is trained on the LVCSR-SI test set and adapted (wherever applicable) using MLLR per-subject adaptation [19]. Test set decoding is performed using a 10.4 K word vocabulary for the LVCSR tasks and the 10-digit vo-

Front End Parameters				Recognition Task		
$m \times n$	D	d	J	LVCSR-SI	LVCSR-SA	DIGITS-SA
64 × 64	24	24	7	105.00	89.19	45.71
64 × 64	50	24	7	98.04	85.42	41.43
64 × 64	50	30	7	97.97	85.50	41.43
64 × 64	100	30	7	96.81	85.84	43.57
64 × 64	50	30	5	99.66	86.42	46.43
64 × 64	50	30	6	99.07	86.63	42.14
64 × 64	50	30	7	97.97	85.50	41.43
64 × 64	50	30	8	96.92	85.31	45.00
64 × 64	50	30	9	96.52	84.86	47.86
64 × 64	50	30	10	94.67	84.81	43.57
64 × 64	50	30	11	94.31	84.09	38.57
64 × 64	100	30	7	96.81	85.84	43.57
80 × 80	100	30	7	95.31	82.98	43.57
96 × 96	100	30	7	94.97	82.65	40.00
112 × 112	100	30	7	93.52	82.51	35.71
128 × 128	100	30	7	93.55	82.65	37.14
64 × 64	24	24	7	105.00	89.19	45.71
128 × 128	100	30	11	91.62	82.31	29.29

Table 3. Visual-only word error rates (%) on the ASR tasks of Table 1, when varying the following visual front end parameters: D , d (first block), J (second block), m , n (third block). The baseline and the proposed visual front end performances are repeated in the last two table lines.

cabulary in the DIGITS-SA task. Similarly to Table 2, the effects of the three visual front end improvements, introduced in Section 3, are studied in separate table blocks, with the baseline and proposed front end performances depicted in the last two lines. For the LVCSR-SI (LVCSR-SA) task, the visual-only WER improves by an absolute 13.4% (6.9%), whereas for the DIGITS-SA task, the performance gain reaches a 36% relative reduction of the WER (from 45.71% to 29.29%).¹ Clearly therefore, the proposed visual front end conveys significantly more speech information than the baseline front end of [9].

5. SUMMARY AND FUTURE WORK

In this paper, we studied three aspects in the design of image (discrete cosine) transform based visual features used for automatic speechreading: (a) The choice of the ROI; (b) The extraction of discriminant image transform features at each frame; and (c) The use of sufficient temporal information to improve visual speech modeling. We demonstrated that the jaw and cheeks do provide useful visual speech information, that can be exploited to improve automatic speechreading using discriminant within-frame feature extraction. Across-frame feature LDA over long temporal windows also improves performance. Combined, all improvements reduced visual-only word error rate by 13% (absolute) in a speaker-independent LVCSR task, and by 36% (relative) in a small, speaker-adapted, connected-digits task.

We are currently investigating the improved visual front end benefit to joint audio-visual LVCSR, by using a number of feature and decision fusion techniques for audio-visual integration, studied in [8]. The phonetic classification experiment, reported in Section 4, where an additional 28% relative improvement in

¹Due to the small DIGITS-SA test set (see Table 1), small WER differences are not significant, and when comparing visual front end parameters, such differences deviate from the trends observed in the LVCSR tasks.

audio-visual over audio-only classification accuracy was observed due to the proposed visual features (as opposed to the baseline visual front end), is indicative that the improved visual-only ASR performance reported in this work should translate into superior audio-visual recognition in the LVCSR domain.

6. REFERENCES

- [1] D.W. Massaro and D.G. Stork, "Speech recognition and sensory integration," *American Scientist*, 86(3):236-244, 1998.
- [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] A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 461-471, 1996.
- [4] A. Rogozan, P. Deléglise, and M. Alissali, "Adaptive determination of audio and visual weights for automatic speech recognition," *Proc. Europ. Tut. Res. Work. Audio-Visual Speech Process.*, pp. 61-64, 1997.
- [5] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, and A. Waibel, "Toward movement-invariant automatic lip-reading and speech recognition," *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 109-112, 1995.
- [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.*, vol. 6, pp. 3733-3736, 1998.
- [7] S. Dupont and J. Luetin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, 2:141-151, 2000.
- [8] C. Neti, G. Potamianos, J. Luetin, I. Matthews, D. Ver-gyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," *Final Summer 2000 Workshop Report*, Center for Language and Speech Processing, Baltimore, 2000 (http://www.clsp.jhu.edu/ws2000/final_reports/avsr/).
- [9] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, "A cascade image transform for speaker independent automatic speechreading," *Proc. Int. Conf. Multimedia Expo.*, vol. II, pp. 1097-1100, 2000.
- [10] I. Matthews, G. Potamianos, C. Neti, and J. Luetin, "A comparison of model and transform-based visual features for audio-visual LVCSR," to appear: *Proc. Int. Conf. Multimedia Expo.*, 2001.
- [11] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," *Proc. Int. Conf. Image Process.*, vol. III, pp. 173-177, 1998.
- [12] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, 264:746-748, 1976.
- [13] A.Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading*, B. Dodd and R. Campbell eds., Lawrence Erlbaum Associates, Hillsdale, pp. 97-113, 1987.
- [14] P.M.T. Smeele, "Psychology of human speechreading," in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 3-15, 1996.
- [15] T.R. Jordan and P.C. Sergeant, "Effects of facial image size on visual and audio-visual speech," in *Hearing by Eye II*, R. Campbell, B. Dodd, and D. Burnham eds., Psychology Press Ltd. Publishers, Hove, pp. 155-176, 1998.
- [16] L.D. Rosenblum and H.M. Saldaña, "Time-varying information for visual speech perception," in *Hearing by Eye II*, R. Campbell, B. Dodd, and D. Burnham eds., Psychology Press Ltd. Publishers, Hove, pp. 61-81, 1998.
- [17] C.R. Rao, *Linear Statistical Inference and Its Applications*. John Wiley and Sons, New York, 1965.
- [18] R.A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 2, pp. 661-664, 1998.
- [19] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech Lang.*, 9:171-185, 1995.