A Low Complexity Parabolic Lip Contour Model With Speaker Normalization For High-Level Feature Extraction in Noise Robust Audio-Visual Speech Recognition

Bengt J. Borgström, *Student Member, IEEE,* and Abeer Alwan, *Senior Member, IEEE*

***Abstract*— This paper proposes a novel low complexity lip contour model for high-level optic feature extraction in noise robust AV-ASR systems. The model is based on weighted least-squares parabolic fitting of the upper and lower lip contours and does not require the assumption of symmetry across the horizontal axis of the mouth, and therefore is realistic. The proposed model does not depend on accurate estimation of specific facial points, as do other high-level models. Also, we present a novel low complexity algorithm for speaker normalization of the optic information stream, which is compatible with the proposed model, and which does not require parameter training. Use of the proposed model with speaker normalization results in noise robustness improvement in isolated audio-visual word recognition relative to using the baseline high-level model.**

I. Introduction

It is well known that human perception of speech relies on both acoustic and visual information [1]. Analogously, the performance of machine recognition of speech has been shown to improve with the presence of optic features along with acoustic features [4]. Optic information in audiovisual automatic speech recognition (AV-ASR) is especially useful when the acoustic signal is degraded or ambiguous. This result has been shown for numerous systems and databases [2]-[7].

The benefit of AV-ASR over acoustic-only recognition is due to the complementary nature of the acoustic and optic information streams within speech [6] [7]. For example, the plosives /*p*/ and /*b*/ are visually indistinguishable, though their acoustic signals can be distinguished due to the voicing of /*b*/. Conversely, the nasal phonemes /*n*/ and /*m*/ can be acoustically similar, though their visual production patterns differ.

Integral components of an AV-ASR system are the methods used for feature extraction and modeling of the acoustic and optic signals. Acoustic feature extraction methods have been researched for many decades, and certain feature vectors, such as Mel Frequency Cepstral Coefficients (MFCCs) and Linear Prediction Cepstral Coefficients (LPCCs), are widely used [8]. However, optic feature extraction and modeling is currently a major topic of research.

The purpose of optic feature extraction in an AV-ASR system is to provide information about the visual aspects of the speakers' speech production to the back-end recognition engine. Various approaches have been proposed for optic feature extraction within AV-ASR systems. Some of these approaches involve image transform based processing of streaming video signals [5], [4]; these methods are referred to as low-level feature extraction methods. The main objective of low-level optic feature extraction is to perform dimension reduction of the raw optic signal, due to the large size of the streaming video, while retaining the majority of discriminative information. Algorithms commonly used for dimension reduction of optic information include Principle Component Analysis (PCA) [5], 2-Dimensional Discrete Cosine Transforms (2-D DCT) [4], and Linear Discriminant Analysis (LDA) [4]. Low-level feature extraction methods are of low complexity and can include information about the tongue and teeth, but are very sensitive to environmental characteristics such as lighting, head rotation, and color.

Other optic feature extraction approaches involve estimation of facial feature information [9] [11], and are referred to as high-level feature extraction methods. The objective of high-level feature extraction is to model facial components that are important to speech recognition, such as the lip contour, and to estimate the parameters of the model. High-level feature extraction methods generally involve high complexity computations, but are robust to many environmental aspects that may lead to poor results for low-level methods.

An important aspect of high-level feature extraction algorithms for AV-ASR systems is the choice of model parameters that comprise the optic signal feature vector. In [7], the authors utilize a parameterized model

of the mouth consisting of height and width of the mouth opening. However, this model assumes static and dynamic symmetry of the mouth across both the horizontal and vertical axes. In [12], the authors introduce a parameterized lip contour model comprised of the mouth width, and the height of opening of the upper and lower lips separately. This model relaxes the constraint of symmetry across the horizontal axis. However, both of these models rely on accurate estimation of specific facial points.

This paper proposes a novel low complexity lip contour model based on weighted least-squares parabolic fitting of the upper and lower lip contours. The model does not require the assumption of static symmetry of the mouth across the horizontal axis, and therefore provides a more reliable model for the lip contour.

Another important aspect of the proposed model is that it is compatible with a variety of lip contour extraction methods, and is robust to missing or noisy data points. That is, calculation of the parameters of our model does not rely on accurate extraction of specific facial points. Instead, construction of the parabolic model simply requires a set of weighted points that may be extracted from arbitrary positions along the lip contour in the optic signal.

Finally, we present a novel low complexity speaker normalization algorithm that is compatible with our proposed model. The algorithm includes no data-dependent parameters, and therefore does not require any training. The proposed speaker normalization technique is therefore applicable to speaker-independent AV-ASR systems, and does not require knowledge of the speaker's identity during testing.

In Section II we discus the role of optic feature extraction in AV-ASR systems and include descriptions of existing high-level feature models. We then introduce the proposed model in Section III. Next, we introduce the proposed speaker normalization technique in Section IV. Section V provides experimental results and analysis of the implemented AV-ASR system using various feature models. Finally, we provide conclusions in Section VI.

II. Optic Information Stream Feature Extraction

A. Low-Level Feature Extraction

The objective of low-level optic feature extraction methods is to reduce the high dimensionality of streaming video while retaining the majority of discriminative information. These methods apply various transforms to individual image frames. The benefit of low-level methods is the low complexity involved. Also, information about the tongue and teeth, which is beneficial to lipreading [1], can be retained. However, low-level methods are very sensitive to environmental differences between speakers, such as lighting and head rotation. Also, accurate and reliable localization of the mouth region is extremely important to low-level extraction algorithms, and thus such techniques may rely on face localization or face tracking [10].

In [5], the authors first normalize image frames for lighting conditions using histogram equalization. Next, PCA is applied to the mouth region of the image frame, and the top 32 coefficients are retained. Additionally, to capture dynamic information, the PCA components of 3 consecutive frames are concatenated to create a 96-dimensional feature vector. This method is sensitive to localization of the mouth region. Also, the high dimensionality of the optic feature vector may lead to inaccurate training of AV-ASR models due to lack of sufficient training data.

B. High-Level Feature Extraction

The objective of high-level optic feature extraction in AV-ASR systems is to extract information about facial features in order to construct a model of the visual speech production system. As can be expected, highly detailed facial models can lead to better performance of AV-ASR systems, as opposed to simple models. However, the construction of such detailed facial models involve highly complex algorithms. In [11], the authors implement an optic feature extraction method that is compliant with the MPEG-4 audiovisual synthesis standard [15]. This method involves mouth localization, a Gradient Vector Flow (GVF) snake algorithm, parabolic lip contour fitting based on the GVF results, and finally extraction of 68 MPEG-4 compliant features. In order to apply these features to an AV-ASR system, dimension reduction is performed through PCA.

The complexity of the previously described feature extraction algorithm may present problems for real-time AV-ASR systems. Instead, we analyze simpler high-level facial models. This paper focuses on the last component of the optic feature extraction system, after initial processing of the video signal and extraction of information regarding the lip contour of the current speaker. The prior stages often use edge detection and/or color discrimination for lip tracking to provide possible lip contour points with corresponding pixel-specific log-likelihoods or weights [12]. In this scenario, the overall objective of simple high-level feature extraction models is to accurately model the speakers' lip contour based on noisy or incomplete information.

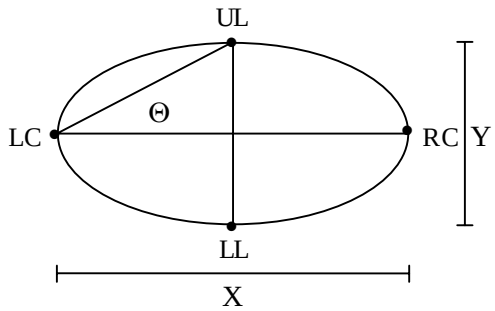

Fig. 1. *The 2-Parameter Lip Contour Model Introduced in [7].*

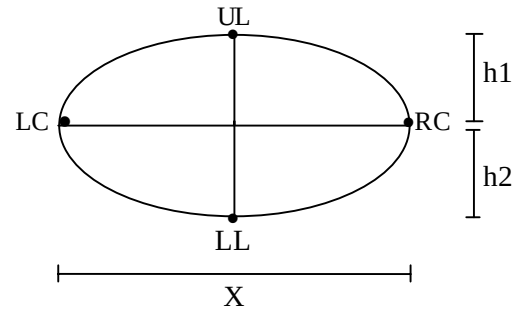

Fig. 2. *The 3-Parameter Lip Contour Model Introduced in [12]. Note that h_1 and h_2 are equal in the figure, but in general this is not required.*

C. The 2-Parameter Γ_2 Model

The authors of [7] introduced a lip contour model represented by an ellipse with its foci along the horizontal axis. The model is defined by the width of the mouth opening, X, and the height of the mouth opening, Y. This model will be referred to as $\Gamma_2 = \{X, Y\}$, since it is defined by 2 parameters. The parameters of Γ_2 are determined by 4 facial positions extracted from the optic signal, namely the left and right corners of the mouth, and the centers of the upper and lower lips. An additional parameter, Θ, is introduced, and is defined as:

$$\Theta = \arctan\left(\frac{Y}{X}\right). \tag{1}$$

The Γ_2 model is shown in Figure 1. Note that LC and RC represent the left and right corner of the mouth respectively, and UL and LL represent the centers of the upper and lower lip, respectively.

As can be interpreted from Figure 1, the Γ_2 model assumes symmetry along both the horizontal and vertical axes, which is rarely the case in humans. Another drawback of the model is that it requires dynamics of the upper and lower lip to be synchronized, which is not generally true in human speech [1]. Finally, the model depends heavily on extraction of the 4 facial feature points shown in Figure 1. However, extraction of such precise points is a complex image processing task, and therefore cannot always be performed with the required accuracy.

D. The 3-Parameter Γ_3 Model

The lip contour model introduced in [12], which will be referred to as the Γ_3 model, includes 3 defining parameters. They are the width of the mouth opening, X, as well as the height of the lower and upper lip openings, h_1 and h_2. Thus, this model relaxes the constraint of symmetry along the horizontal axis that is present in the Γ_2 model. The 3-parameter Γ_3 model is illustrated in Figure 2.

It can be concluded from Figure 2 that the Γ_3 model includes separate parameters describing the shape and motion of the upper and lower lips. Thus, it provides a more accurate model of the mouth during speech. However, the Γ_3 model still relies on the accurate extraction of the 4 specific facial positions discussed in Section II-C.

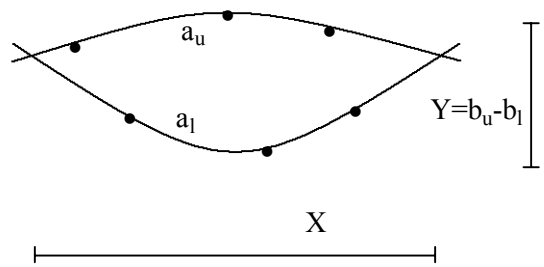

Fig. 3. *The Proposed Parabolic Lip Contour Model*

III. The Parabolic Lip Contour Model

A. Description of the Proposed Γ_P Model

We propose a parameterized lip contour model , Γ_P, based on a pair of intersecting parabolas with opposite orientation. The defining parameters of the model include the focal parameters of the upper and lower parabolas, a_u and a_l, respectively, and X and Y, the difference between the offset parameters of the parabolas, b_u and b_l. Note that the focal parameters can be interpreted as measures of rounding of the lips. The proposed model is shown in Figure 3.

The parabolic model includes separate parameters for the motion of the upper and lower lips of the mouth during speech. Thus, it relaxes the constraint of symmetry present in the Γ_2 model.

Another benefit of the proposed model is that it does not rely on lip contour points extracted from specific facial positions, as do the Γ_2 and Γ_3 models previously described. Instead, the values a_u, a_l, and Y depend on any set of arbitrarily collected lip contour points and their corresponding weights. That is, the parabolic model can be constructed using a group of lip contour points that need not be extracted from specific facial locations, such as the corners of the mouth or the centers of the lips. Conversely, if a scenario exists in which any subset of the points $\{LC, RC, UL, LL\}$ can not be estimated accurately, the models Γ_2 and Γ_3 may fail.

B. Derivation of the Γ_P Parameters

Let us assume that sets of lip contour points have been extracted from the upper lip and lower lip of the raw optic signal, and let these sets be represented by

$$S_U = \{(x_u(1), y_u(1)), \ldots, (x_u(N_U), y_u(N_U))\} \quad (2)$$

and

$$S_L = \{(x_l(1), y_l(1)), \ldots, (x_l(N_L), y_l(N_L))\}, \quad (3)$$

where N_U and N_L represent the number of points extracted from the upper and lower lips respectively. Also, assume that the weighting vectors $\mathbf{w}_u$ and $\mathbf{w}_l$ have been determined, where $w_u(i)$ and $w_l(i)$ represent the weights or reliability measures of the ith points in the sets S_U and S_L, respectively. The actual calculation of $\mathbf{w}_u$ and $\mathbf{w}_l$ is dependent on the specific algorithm used to extract the points comprising S_U and S_L from the raw video, and thus falls outside the scope of this paper. However, the proposed model offers the framework to use reliability measures of individual points from the upper and lower lips, if applicable. Note that the database used in this study [12] does not supply individual weights, and thus all points are equally weighted so that $\mathbf{w}_u$=$\mathbf{q}_u$ and $\mathbf{w}_l$=$\mathbf{q}_l$, where:

$$\mathbf{q}_u = [1, 1, \ldots, 1]^T_{1\times N_U}, \quad (4)$$

and

$$\mathbf{q}_l = [1, 1, \ldots, 1]^T_{1\times N_L}.$$

Let us define the vectors:

$$\mathbf{x}_u = [x_u(1), \ldots, x_u(N_u)]^T, \quad (5)$$

$$\mathbf{y}_u = [y_u(1), \ldots, y_u(N_u)]^T, \quad (6)$$

$$\mathbf{x}_l = [x_l(1), \ldots, x_l(N_L)]^T, \text{ and} \quad (7)$$

$$\mathbf{y}_l = [y_l(1), \ldots, y_l(N_L)]^T. \quad (8)$$

An estimated midpoint of the mouth along the horizontal axis, $\widetilde{x}_c$, can be found as:

$$\widetilde{x}_c = \frac{\mathbf{w}_u^T\mathbf{x}_u + \mathbf{w}_l^T\mathbf{x}_l}{\mathbf{w}_u^T\mathbf{q}_u + \mathbf{w}_l^T\mathbf{q}_l}. \quad (9)$$

Thus, we can determine the normalized vectors:

$$\widehat{\mathbf{x}}_u = [\widehat{x}_u(1), \ldots, \widehat{x}_u(N_u)]^T, \text{ and} \quad (10)$$
$$\widehat{\mathbf{x}}_l = [\widehat{x}_l(1), \ldots, \widehat{x}_l(N_L)]^T,$$

where $\widehat{x}_u(i) = x_u(i) - \widetilde{x}_c$ and $\widehat{x}_l(i) = x_l(i) - \widetilde{x}_c$. Now let us define the diagonal matrices:

$$\mathbf{M}_U = \begin{bmatrix} \widehat{x}_u(1) & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \widehat{x}_u(N_U) \end{bmatrix}_{N_U\times N_U}, \quad (11)$$

and:

$$\mathbf{M}_L = \begin{bmatrix} \widehat{x}_l(1) & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \widehat{x}_l(N_L) \end{bmatrix}_{N_L\times N_L}. \quad (12)$$

Additionally, define the diagonal matrices:

$$\mathbf{W}_U = \begin{bmatrix} w_u(1) & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & w_u(N_U) \end{bmatrix}_{N_U\times N_U}, \quad (13)$$

and:

$$\mathbf{W}_L = \begin{bmatrix} w_l(1) & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & w(N_L) \end{bmatrix}_{N_L\times N_L}. \quad (14)$$

Consider then the vector-form parabolic functions:

$$\mathbf{p}_u = a_u\mathbf{M}_U\widehat{\mathbf{x}}_u + b_u\mathbf{q}_u, \text{ and} \quad (15)$$
$$\mathbf{p}_l = a_l\mathbf{M}_L\widehat{\mathbf{x}}_l + b_l\mathbf{q}_l, \quad (16)$$

where a_u and a_l are the focal parameters of the parabolas, and b_u and b_l are the offset constants. Additionally, $\mathbf{Q}_U$ and $\mathbf{Q}_L$ are defined as:

$$\mathbf{Q}_U = \begin{bmatrix} 1 \cdots 1 \\ \vdots \ddots \vdots \\ 1 \cdots 1 \end{bmatrix}_{N_U \times N_U}, \text{ and } \mathbf{Q}_L = \begin{bmatrix} 1 \cdots 1 \\ \vdots \ddots \vdots \\ 1 \cdots 1 \end{bmatrix}_{N_L \times N_L}. \tag{17}$$

The weighted least-squares parabolic fit of the set of lip contour points extracted from the upper lip, S_U, can thus be determined by minimizing the least-squares cost function:

$$J_U = |\mathbf{W}_U (\mathbf{p}_u - \mathbf{y}_u)|^2 = |\mathbf{W}_U (a_u \mathbf{M}_U \widehat{\mathbf{x}}_u + bu \mathbf{q}_u - \mathbf{y}_u)|^2. \tag{18}$$

Taking the partial derivative of the cost function J_U with respect to the parabolic focal parameter a_u, and equating this expression to 0, results in the following expression:

$$\begin{aligned} \frac{\partial J_U}{\partial a_u} =& a_u \widehat{\mathbf{x}}_u^T \mathbf{W}_U^2 \mathbf{M}_U^2 \widehat{\mathbf{x}}_u + b_u \mathbf{q}_u^T \mathbf{W}_U^2 \mathbf{M}_U \widehat{\mathbf{x}}_u \\ & - \mathbf{y}_u^T \mathbf{W}_U^2 \mathbf{M}_U \widehat{\mathbf{x}}_u = 0. \end{aligned} \tag{19}$$

Similarly, taking the partial derivative of the cost function J_U with respect to the parabolic offset parameter b_u, and equating the resulting derivative to 0, leads to the following expression:

$$\frac{\partial J_U}{\partial b_u} = b_u \mathbf{q}_u^T \mathbf{W}_U^2 \mathbf{q}_u - \mathbf{q}_u^T \mathbf{W}_U^2 \mathbf{y}_u + a_u \mathbf{q}_u^T \mathbf{W}_U^2 \mathbf{M}_U \widehat{\mathbf{x}}_u = 0. \tag{20}$$

Substitution of Equation 20 into Equation 19 results in an expression for the parabolic focal parameter a_u:

$$a_u = \frac{\mathbf{y}_u^T \mathbf{H}_U \mathbf{M}_U \mathbf{x}_u}{\mathbf{x}_u^T \mathbf{M}_U \mathbf{H}_U \mathbf{M}_U \mathbf{x}_u}, \tag{21}$$

where:

$$\mathbf{H}_U = \left(\mathbf{q}_u^T \mathbf{W}_U^2 \mathbf{q}_u \mathbf{I}_{N_U} - \mathbf{W}_U^2 \mathbf{Q}_U\right) \mathbf{W}_U^2. \tag{22}$$

An expression for the offset parameter b_u can then be obtained by solving Equation 20, which results in:

$$b_u = \frac{\left(\mathbf{y}_u^T - a_u \mathbf{x}_u^T \mathbf{M}_U\right) \mathbf{W}_U^2 \mathbf{q}_u}{\mathbf{q}_u^T \mathbf{W}_U^2 \mathbf{q}_u}. \tag{23}$$

Expressions for the parameters, a_l and b_l, of the least-squares parabolic fit to the lower lip contour can be found by applying the corresponding steps described by Equations 18-23 to $J_L = |\mathbf{W}_L (\mathbf{p}_l - \mathbf{y}_l)|^2$, leading to:

$$a_l = \frac{\mathbf{y}_l^T \mathbf{H}_L \mathbf{M}_L \mathbf{x}_l}{\mathbf{x}_l^T \mathbf{M}_L \mathbf{H}_L \mathbf{M}_L \mathbf{x}_l}, \tag{24}$$

where:

$$\mathbf{H}_L = \left(\mathbf{q}_l^T \mathbf{W}_L^2 \mathbf{q}_l \mathbf{I}_{N_L} - \mathbf{W}_L^2 \mathbf{Q}_L\right) \mathbf{W}_L^2, \tag{25}$$

and:

$$b_l = \frac{\left(\mathbf{y}_l^T - a_l \mathbf{x}_l^T \mathbf{M}_L\right) \mathbf{W}_L^2 \mathbf{q}_l}{\mathbf{q}_l^T \mathbf{W}_L^2 \mathbf{q}_l}. \tag{26}$$

The previously mentioned lip contour models Γ_2 and Γ_3 can be derived from the parabolic parameters given in Equations 21, 23, 24, and 26. Specifically, the models can be approximated as:

$$Y = b_u - b_l, \tag{27}$$

$$X = 2\sqrt{\frac{b_l - b_u}{a_u - a_l}}, \tag{28}$$

$$\Theta = \arctan\left\{\frac{\sqrt{(a_l - a_u)(b_u - b_l)}}{2}\right\}, \tag{29}$$

$$h_1 = a_u \left(\frac{b_l - b_u}{a_l - a_u}\right), \tag{30}$$

and:

$$h_2 = a_l \left(\frac{b_u - b_l}{a_l - a_u}\right). \tag{31}$$

Thus, our proposed parabolic lip contour model can be constructed from the sets of extracted points S_U and S_L, and the corresponding weighting vectors $\mathbf{w}_u$ and $\mathbf{w}_l$. As stated previously, these sets of lip contour points need not contain information about specific facial positions such as the corner of the mouth and center of the lips. Instead, the sets can include only those extracted lip contour points deemed reliable. Also, it can be expected that the accuracy of the Γ_P model will increase as the number of elements in S_U and S_L increases.

IV. Speaker Normalization

Speaker normalization for acoustic-only ASR has been shown to be an effective method to counter the performance degradation caused by mismatch between speakers with respect to physical characteristics of speech production [16]. Speaker normalization applies transformations in the feature domain with the aim of warping the frequency scale to align power spectra prior to the recognition process. Variability in speech acoustics is often caused by age and/or gender differences in speakers [17].

Analogously, speaker normalization can be applied to optic features to counter the effect of speaker differences. During modeling of facial features in high-level optic

TABLE I

Mean and Variance Values of Y Averaged Over 10 Takes For Each Speaker, Obtained on the AV-ASR Database From [12]

Speaker	**Mean** (μ)	**Variance** (σ^2)
1	50.25	50.13
2	49.79	121.76
3	64.64	151.48
4	44.50	68.04
5	54.44	113.76
6	62.14	192.09
7	63.78	331.83
8	51.38	61.92
9	48.25	57.92
10	72.29	199.45

feature extraction within an AV-ASR system, there exist scenarios in which model parameters corresponding to similar utterances may be of different magnitudes for various speakers. These situations include comparing speakers that have differing face sizes and differing speech production characteristics, as well as comparing raw optic data obtained with varying distances between the speaker and the camera. Table I shows the mean and variance values of the parabolic focal parameter Y for each of 10 speakers averaged over 10 takes. Note the wide range in the Y parameter values for different speakers.

In order to better recognize patterns of visual speech production within the speaker independent AV-ASR back engine, optic features can be normalized. That is, the back end recognizer can be expected to provide improved results if the optic features corresponding to the same utterances lay within the same approximate range for each speaker.

An intuitive approach to normalization of high-level parameter values is to view each parameter in time as a stochastic process with normal distribution, defined by the pair $\left(\mu, \sigma^2\right)$. If these parameters can be estimated, then the model Γ_k can be normalized, and will be referred to as Γ_k^N.

Let the model Γ_k be comprised of the feature vector $[f_1, f_2, \ldots, f_k]$, where $f_i(n)$, for $1 \leq n \leq M$, contains the given parameter values in time, and where M is the length of the utterance in samples. The normalized feature vector f_i^N can then be determined by:

$$f_i^N(n) = \frac{(f_i(n) - \mu_i)}{\sigma_i}, \text{ for } 1 \leq n \leq M. \tag{32}$$

The pair $\left(\mu_i, \sigma_i^2\right)$ which parameterizes the process f_i can be approximated based on M frames of the current utterance:

$$\mu_i = \frac{1}{M} \sum_{n=1}^{M} f_i(n), \tag{33}$$

and:

$$\sigma_i^2 = \frac{1}{M} \sum_{n=1}^{M} (f_i(n) - \mu_i)^2. \tag{34}$$

Thus, the optic features for an arbitrary model Γ_k can be normalized without predetermined feature distribution parameters. Note that this process is similar to mean and variance normalization (MVN) in ASR [18], which normalizes acoustic speech features in the cepstral domain to account for additive noise.

V. Experimental Set-up and Results

A. Description of the AV-ASR System and Database

An AV-ASR system was implemented to test the models previously described. The recognition system used early-integration of the optic and acoustic feature vectors, which fuses the feature vectors prior to the recognition process. The back end of the AV-ASR system consisted of a 6-state Hidden Markov Model (HMM), with each of the 4 emitting states including 4 Gaussian mixtures. The HMM back end utilized code from the HTK toolkit.

The developed AV-ASR system was tested using the audio-visual isolated word database from [12]. The database consists of 10 speakers each saying a series of words, and repeating the series 10 times. The raw audio data was in the form of PCM coded signals sampled at $44.1kHz$. The optic data was comprised of the horizontal and vertical positions of the left and right corners of the mouth as well as the heights of the openings of the upper and lower lips, and the optic data was sampled at $30\,Hz$. We used a subset of 10 words from the database, namely the digits from one to ten.

The acoustic feature vector consisted of the first 12 Mel-Frequency Cepstral Coefficients, along with the log-energy of the spectrum. The approximated derivatives and double-derivatives were also included. In order to compare performance improvements provided by AV-ASR to traditional noise robust ASR, we also implemented a noise robust front end utilizing Perceptual Linear Prediction and relative spectral (RASTA) processing, as described in [19] and [20]. Code for implementing RASTA-PLP feature extraction was obtained from [21].

In order to use early-integration of the audio and optic information streams, the optic features were upsampled and linearly interpolated. The upsampled optic

TABLE II

Optic-Only Recognition Results for Individual Features, Obtained on the AV-ASR Database From [12]: f_i represents the recognition rate using the corresponding feature without speaker normalization, and f_i^N represents the recognition rate for the feature with speaker normalization.

Parameter	Rec. Rate for f_i	Rec. Rate for f_i^N
X	13.58%	31.42 %
Y	23.73%	43.24 %
Θ	22.04%	37.09 %
h_1	13.08%	18.12 %
h_2	27.19%	39.38 %
a_u	11.72%	17.52 %
a_l	25.68%	34.73 %

signals were then low-pass filtered to eliminate high-frequency effects. Thus, the audio and optic signals could be windowed with equivalent window lengths, and fused easily. The resulting feature vectors thus included both the acoustic and optic features in a single vector. For example, for AV-ASR using MFCCs including the approximated derivatives and double-derivatives, which results in a 39-element acoustic vector, and using the Γ_3 model, which results in a 3-element optic vector, the final AV-ASR observation vector is comprised of 42 elements.

Due to the relatively limited amount of data in the database, a bootstrapping technique was used to test the AV-ASR system. Each word was included in each of 10 takes, by each of 10 speakers. Therefore, we trained the system using 9 of 10 takes, and tested on the excluded take. This process was repeated 10 times in order to test all data files. The final performance was obtained by averaging the recognition rates over the 10 takes.

B. Optic-Only Recognition Performance of Individual Model Parameters

The AV-ASR system described in Section V-A was tested using each of the individual parameters included in the Γ_2, Γ_3, and Γ_P lip contour models as a 1-dimensional optic feature vector. The word recognition accuracy was recorded for each parameter, and the results are shown in Table II.

As can be concluded from Table II, there is a large discrepancy between the performance of the various individual parameters in optic-only word recognition. Parameters supplying information regarding the vertical opening of the lips provide better performance than those parameters supplying information about the horizontal opening. Additionally, parameters specifically providing information about the lower lip provide better performance than those parameters supplying information about the upper lip. Finally, it can be concluded that the proposed speaker normalization technique improves the performance for each of the features.

TABLE III

Optic-Only Recognition Results for Model Feature Vectors, Obtained using the AV-ASR Database From [12]

Model	Feature Vector	Recognition Rate
Γ_2	$\{X, Y, \Theta\}$	26.98 %
Γ_3	$\{X, h_1, h_2\}$	34.04 %
Γ_P	$\{Y, X, a_u, a_l\}$	37.28 %
Γ_2^N	$\{X^N, Y^N, \Theta^N\}$	57.27 %
Γ_3^N	$\{X^N, h_1^N, h_2^N\}$	59.29 %
Γ_P^N	$\{Y^N, X^N, a_u^N, a_l^N, \Theta^N\}$	61.17 %

C. Optic-Only Recognition Performance of Model Feature Vectors

The AV-ASR system described in Section V-A was then tested using each of the feature vectors introduced by the Γ_2, Γ_3, and Γ_P lip contour models, as well as their normalized versions. The optic-only word recognition accuracy was recorded for each feature vector, and the results are shown in Table III.

As can be concluded from Table III, the Γ_P model provides superior word recognition results to both the Γ_3 and the Γ_2 models. Additionally, it can be concluded that the proposed speaker normalization technique improves the recognition rate of the optic-only system for each high-level model discussed. The best recognition performance was reported for the normalized parabolic model, Γ_P^N.

D. Performance of AV-ASR Systems with Simple High-Level Feature Models and Speaker Normalization in Noise

The overall AV-ASR system was tested across a range of SNR values of the input acoustic signal, and across a range of weights for the optic and acoustic streams within the back-end recognition engine. These information stream weights in an AV-ASR system control the amount of weight placed on either the acoustic or optic observation features during the recognition process [4]. The system was tested using the three high-level lip contour models discussed in Sections II-C, II-D, and III, as well as the normalized versions of these models. The resulting performance of the system with each of the models is shown in Table IV. Note that the SNR value is calculated with the acoustic signal power being the square of the peak clean signal measure, and the noise

TABLE IV

Word Recognition Rate in AWGN, Obtained on the AV-ASR Database From [12]: ASR refers to acoustic-only recognition, using MFCC or RASTA-PLP features. Γ_2 and Γ_3 refer to AV-ASR, using the baseline high-level optic models from [7] and [12], respectively. Γ_P is the proposed parabolic lip contour model. Γ_2^N, Γ_3^N, and Γ_P^N refer to the normalized versions of the previously mentioned models (described in Section IV). Note that AV-ASR uses MFCC features for the acoustic stream.

SNR (dB)	0	5	10	15	20	25	30	35	40
ASR (*MFCC*)	9.24%	13.44%	28.96%	40.57%	58.44%	76.42%	89.26%	93.82%	96.03%
ASR (*RASTA-PLP*)	11.14%	24.15%	38.79%	58.97%	75.15%	86.43%	92.30%	95.13%	95.74%
Γ_2	31.48%	37.45%	47.13%	54.66%	69.60%	84.34%	93.28%	95.65%	95.80%
Γ_3	37.57%	43.85%	51.74%	59.36%	75.24%	87.58%	93.61%	96.15%	96.10%
Γ_P	39.66%	43.67%	51.84%	62.16%	76.38%	88.03%	94.27%	96.03%	96.86%
Γ_2^N	60.45%	64.97%	71.30%	80.66%	87.35%	92.65%	95.74%	97.19%	97.77%
Γ_3^N	59.29%	66.36%	73.00%	80.84%	88.06%	92.61%	95.29%	97.05%	97.50%
Γ_P^N	61.17%	66.57%	74.47%	81.06%	88.76%	93.22%	96.34%	97.05%	97.48%

power being the mean square value of additive white Guassian noise.

For the results obtained in Table IV, word recognition tests were performed for each SNR level at acoustic stream weights within the range $\lambda_a \in [0, 1]$, and at increments of 0.1. Thereafter, the maximum performance was chosen for each level of SNR from among these word recognition rates. A detailed discussion of the effect of stream weights on AV-ASR performance can be found in [4].

In Figure 4, the word recognition rates of the AV isolated word recognition system is shown graphically for the proposed model and speaker normalization technique. In comparison, the baseline systems of acoustic-only ASR, and AV-ASR using the Γ_2 model, are also plotted. Note that the results shown for the Γ_2 model are similar to those given in [12].

As can be interpreted from Table IV, the Γ_P model provides improvement over both the Γ_2 model and the Γ_3 model, with the improvement over the Γ_2 model being greater. Additionally, the speaker normalization technique provides improved word recognition when applied to each of the high-level models discussed. The best overall performance was obtained using the normalized parabolic model. The Γ_P^N model can achieve a recognition rate of 93.22% at SNR=25 dB, while the baseline Γ_2 model achieves a similar rate at approximately SNR=30 dB. The baseline acoustic-only ASR system achieves similar performance at approximately SNR=35 dB. In comparison with traditional noise robust ASR using PLPCCs and RASTA processing, the proposed AV-ASR systems provides superior performance, especially for low levels of speech signal SNR.

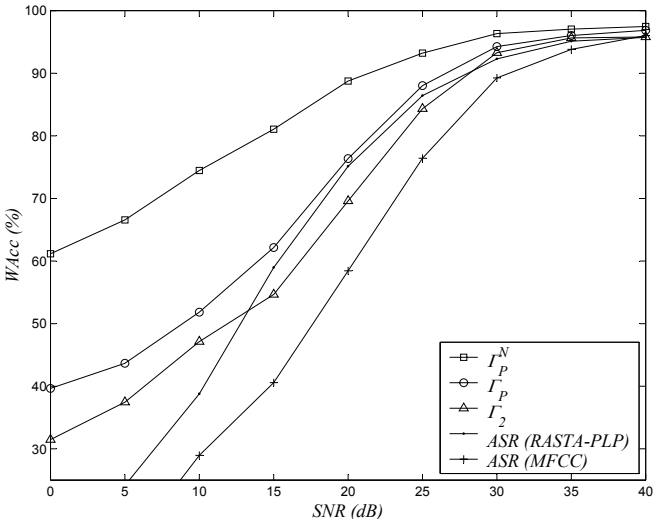

Fig. 4. *Word Recognition Rate of the AV-ASR System Operating With Various High-Level Models. For details, see Table IV Caption.*

VI. Conclusion

This paper focuses on low complexity lip contour models used for high-level optic feature extraction in noise robust AV-ASR systems. The proposed model, Γ_P, is shown to provide improved audio-visual isolated word recognition relative to the lip contour models introduced in [7] and [12]. Additionally, the proposed model does not depend on accurate estimation of specific facial points, as do the Γ_2 and Γ_3 models. Thus, the proposed model is applicable in the scenario of missing or noisy data, when other models may fail.

Additionally, this paper introduces a low complexity speaker normalization technique that requires no training. The speaker normalization technique is shown to provide improved performance when applied to each of the high-level feature models discussed. The best overall

performance of the AV-ASR system was obtained using the normalized parabolic model, Γ_P^N.

Future work includes analyzing the proposed model and speaker normalization method within a more realistic continuous AV-ASR system. Such studies are currently limited by the lack of publicly available AV-ASR databases.

VII. Acknowledgments

This work was supported in part by the NSF.

References

[1] H. McGurk, and J. MacDonald, *Hearing Lips and Seeing Voices*, Nature, vol. 264, no. 5588, pp.746-748, 1976.

[2] G. Papandreou, A. Katsamanis, P. Athanassios, and P. Maragos, *Multimodal Fusion and Learning with Uncertain Features Applied to Audiovisual Speech Recognition*, IEEE Workshop on Multimedia Signal Processing, pp. 264-267, 2007.

[3] M. Hasegawa-Johnson, *A Multi-Stream Approach to Audiovisual Automatic Speech Recognition*, IEEE Workshop on Multimedia Signal Processing, pp. 328-331, 2007.

[4] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, *Dynamic Bayesian Networks for Audio-Visual Speech Recognition*, EURASIP Journal on Applied Signal Processing, vol. 11, pp. 1-15, 2002.

[5] T. J. Hazen, *Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition*, IEEE Trans. on Speech and Audio Processing, vol. 14, no. 3, pp. 1082-1089, 2006.

[6] T. Chen, and R. Rao, *Audio-Visual Integration in Multimodal Communication*, Proc. IEEE, vol. 86, pp.837-852, 1998.

[7] M. N. Kaynak, Q. Zhi, A. D. Cheok, K. Sengupta, Z. Jian, and K. C. Chung, *Analysis of Lip Geometric Features for Audio-Visual Speech Recognition* IEEE Trans. Systems, Man, and Cybernetics- Part A, vol. 34, no. 4, pp. 564-570, 2004.

[8] L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[9] P. S. Aleksic, A. K. Katsaggelos, *Comparison of Low- and High-Level Visual Features for Audio-Visual Continuous Automatic Speech Recognition*, ICASSP, pp. 917-920, 2004.

[10] D. Nguyen, D. Halupka, P. Aarabi, and A. Sheikholeslami, *Real Time Face-Localization Using Field Programmable Gate Arrays*, IEEE Trans. on Systems, Man, and Cybernetics: Part B, vol. 36, no. 4, pp. 902-912, 2006.

[11] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, *Audio-Visual Speech Recognition Using MPEG-4 Compliant Visual Features*, EURASIP Journal on Applied Signal Processing vol. 11, pp. 1213-1227, 2002.

[12] Advanced Multimedia Processing Lab, http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/, Carnegie Mellon University, Pittsburgh, Pa, USA.

[13] M. T. Chan, Y. Zhang, and T. S. Huang, *Real-Time Lip Tracking and Bimodal Continuous Speech Recognition*, in Proc. IEEE 2nd Workshop Multimedia Signal Processing, Los Angeles, CA, pp. 65-70, 1998.

[14] The Open AVCSR Toolkit (April, 2003) [Online] Available from http://sourceforge.net/projects/opencvlibrary/

[15] Text for ISO/IEC FDIS 14496-1 Systems, ISO/IEC JTC1/SC29/WG11 N2502, November 1998.

[16] E. B. Gouvea and R. M. Stern, *Speaker Normalization through Formant-Based Warping of the Frequency Scale*, Eurospeech, vol. 3, pp.1139-1142, 1997.

[17] X. Cui and A. Alwan, *Adaptation of Children's Speech with Limited Data Based on Formant-like Peak Alignment*, Computer Speech and Language, Vol. 20, Issue 4, pp. 400-419, October 2006.

[18] O. Viikki and K. Laurila, *Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition*, Speech Communication, vol. 25, pp. 133-147, 1998.

[19] H. Hermansky, *Perceptual Linear Predictive (PLP) Analysis of Speech*, Journal of the Acoustic Society of America, 87(4), pp. 1738-1752, April 1990.

[20] H.Hermansky and N. Morgan, *RASTA Processing of Speech*, IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 4, pp. 578-589, Oct. 1994.

[21] http://labrosa.ee.columbia.edu/matlab/rastamat/