Speaker and Speech recognition by Audio-Visual lip biometrics

Maycel Isaac Faraj and Josef Bigun

Halmstad University, School of Information Science,
Computer and Electrical Engineering (IDE)
Halmstad University, Box 823, SE-301 18 Halmstad
{maycel.faraj, josef.bigun}@ide.hh.se

Abstract. This paper proposes a new robust bi-modal audio visual speech and speaker recognition system by lip-motion and speech biometrics. To increase the robustness of speech and speaker recognition, we have proposed a method using speaker lip motion information extracted from video sequences with low resolution (128 ×128 pixels). In this paper we investigate a biometric system for speech recognition and speaker identification based using line-motion estimation with speech information and Support Vector Machines. The acoustic and visual features are fused at the feature level showing favourable results with digit recognition being 83% to 100% and speaker recognition 100% on the XM2VTS database.

1 Introduction

In recent years, some techniques have been suggested that combine visual features to improve the recognition rate in acoustically noisy environments that have background noise or cross talk among speakers [1][2][3][4][5]. The present work is a continuation of [6]. The dynamic visual features are suggested based on the shape and intensity of the lip region [7][8][9][10][11] because changes in the mouth shape including the lips and tongue carry significant phoneme-discrimination information. So far the visual representation has been based on shape models to represent changed mouth shapes that rely exclusively on the accurate detection of the lip contours, often a challenging task under varying illumination conditions and rotations of the face. Another disadvantage is the fluctuating computation time due to the iterative convergence process of the contour extraction. The motion in dynamic lip images can be modelled by moving-line patterns also known as *normal image velocity* [6][12].

Here we use direct feature fusion to obtain the audio-visual observation vectors by concatenating the audio and visual features. The observation sequences are then modelled with a Support Vector Machine (SVM) classifier for speech and speaker recognition respectively. The studies [13][14] [15] reported good performance with Support Vector Machine (SVMs) classifiers in recognition, whereas traditional methods for speaker recognition are GMMs [16] and artificial neural networks [17]. By investigating SVM instead of the more common GMM [6], we wanted to study the performance influence of the classification method on speaker recognition and speech recognition.

Here, we extended previous work [6] by studying novel quantization technique for lip features in two additional applications - audio-visual speech recognition and biometric speaker identification. The reminder of the paper is organized as follows. In Section 2 we describe briefly the lip-motion technique for the mouth region along with our quantization (feature-reduction) method, followed by acoustic feature extraction in Section 3. In Section 4, the database used and the experimental setup are described. Section 5 describes SVM classifiers used for speech and speaker recognition, and the experimental results are shown in Section 6. Finally, we conclude with a discussion of the experiments and the remaining issues.

2 Visual features by normal image velocity

Bigun et al. proposed a different motion estimation technique based on an eigenvalue analysis of the multidimensional structure tensor [18], allowing the minimization process of fitting a line or a plane to be carried without the Fourier Transform. Applied to optical-flow estimation, known as the 3D structure-tensor method, the eigenvector belonging to the largest eigenvalue of the tensor is directed in the direction of the contour motion, if motion is present. However, this method can be excessive for applications that need only line-motion features. We assume that the local neighbourhood in the lip image contains parallel lines or edges, this assumption is realistic [19]. Lines in spatio-temporal image translated with a certain velocity in the normal direction will generate planes with a normal that can be estimated in a total-least-square-error (TLS) sense as the local directions of the lines in 2D manifolds using complex arithmetic and convolution [18]. The velocity component of translation parallel to the line cannot be calculated; this is referred to as the *aperture problem.* We denote the normal unit vector as $\mathbf{k} = (k_x, k_y, k_t)^T$ and the projection of $\mathbf{k}$ to the x–y coordinate axes represents the direction vector of the line's motion. The normal, $\mathbf{k}$, of the plane will then relate to the velocity vector $v\mathbf{a}$ as follows

$$\mathbf{v} = v\mathbf{a} = -\frac{k_t}{k_x^2+k_y^2}\left(k_x, k_y\right)^T =$$

$$-\frac{1}{(\frac{k_x}{k_t})^2 + (\frac{k_y}{k_t})^2}\left(\frac{k_x}{k_t}, \frac{k_y}{k_t}\right)^T , \qquad (1)$$

where $\mathbf{v}$ is the *normal image flow.* The normal velocity estimation problem becomes a problem of solving the tilts ($\tan\gamma_1 = \frac{k_x}{k_t}$) and ($\tan\gamma_2 = \frac{k_y}{k_t}$) of the motion plane in the xt and yt manifolds, which is obtained from the eigenvalue analysis of the 2D structure tensor, [18]. Using complex numbers and smoothing, the angles of the eigenvectors are given effectively as complex values such that its magnitude is the difference of the eigenvalues of the local structure tensor in the xt manifold, whereas its argument is twice the angle of the most significant eigenvector approximating $2\gamma_1$. The function f represents the continuous

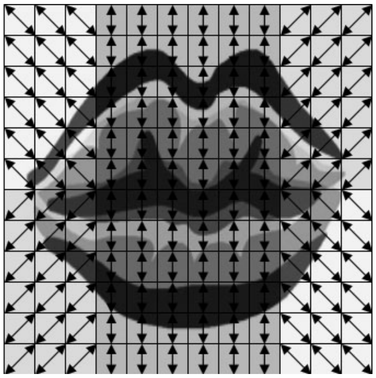

Fig. 1. Illustration of velocity estimation quantification and reduction.

local image, whose sampled version can be obtained from the observed image sequence. Thus, the arguments of $\tilde{u}_1$ and $\tilde{u}_2$ deliver the TLS estimations of γ_1 and γ_2 in the local 2D manifolds *xt* and *yt* respectively, but in the double angle representation [20], leading to the estimated velocity components as follows.

$$\frac{k_x}{k_t} = \tan\gamma_1 = \tan(\frac{1}{2}\arg(\tilde{u}_1)) \Rightarrow \tilde{v}_x = \frac{\tan\gamma_1}{\tan^2\gamma_1 + \tan^2\gamma_2} \tag{2}$$

$$\frac{k_y}{k_t} = \tan\gamma_2 = \tan(\frac{1}{2}\arg(\tilde{u}_2)) \Rightarrow \tilde{v}_y = \frac{\tan\gamma_2}{\tan^2\gamma_1 + \tan^2\gamma_2} \tag{3}$$

The tilde over v_x and v_y donate that these quantities are estimations of v_x and v_y. With the calculated 2D-velocity feature vectors, $(v_x, v_y)^T$, in each mouth-region frame (128×128 pixels) we have dense 2D-velocity vectors. To extract statistical features from the 2D normal velocity and to reduce the amount of data without degrading identity-specific information excessively, we reduce the 2D velocity feature vectors $(v_x, v_y)^T$ at each pixel to 1D scalars where the expected directions of motion are 0°, 45°, −45° – marked with 3 different greyscale shades in 6 regions in **Fig. 1** . The motion vectors within each region become real scalars that take the signs + or − depending on which direction they move relative to their expected spatial directions (differently shaded boxes).

$$f(p,q) = \|(v_x(p,q), v_y(p,q))\| * sgn(\angle(v_x(p,q), v_y(p,q))),\ p,q = 0\ldots127. \tag{4}$$

The next step is to quantize the estimated velocities from arbitrary real scalars to a more limited set of values. Empirically we found that direction and speed quantization are significant reduces the impact of noise on the motion information around the lip area. The quantized speeds are obtained from the data by applying a mean approximation as follows.

$$g(l,k) = \sum_{p,q=0}^{N} f(Ml+q, Nk+p), \;\; p,q = 0\ldots(N-1), \; l,k = 0\ldots(M-1) \quad (5)$$

where N and M represent the window size of the boxes (**Fig. 1**) and the number of boxes, respectively. The statistics of lip-motion are represented by 144-dimensional ($M \times M$) feature vectors. The original dimension before reduction is $128 \times 128 \times 2 = 32768$.

3 Acoustic features

The Mel-Frequency Cepstral Coefficient (MFCC) is a commonly used instance of the filter-bank–based features [21] that can represent the speech spectrum. Here, the input signal is pre-emphasized and divided into 25-ms frame every 10 ms. A Hamming window is applied to each frame that is computed by (MFCC) vectors from the FFT-based, mel-warped, log-amplitude filter bank followed by a cosine transform and cepstral filtering. The speech features in this study were the MFCC vectors generated by the Hidden Markov Model Toolkit (HTK) [22] processing the data stream from the XM2VTS database. This MFCC vector contains 12 cepstral coefficients extracted from the Mel-frequency spectrum of the frame with normalized log energy, 13 delta coefficients (velocity), and 13 delta-delta coefficients (acceleration).

4 XM2VTS database

All experiments in this paper are conducted by the XM2VTS database, currently the largest publicly available audio-visual database [23]. The XM2VTS database contains images and speech of 295 subjects (male and female), captured over 4 sessions. In each session, the subject is asked to pronounce three sentences when recording the video sequence; we use only “0 1 2 3 4 5 6 7 8 9”. It is worth noting that the XM2VTS data is difficult to use as is for speech recognition experiments because the speech or lip motions are not annotated. Before defining a protocol we thus needed to annotate both speech and visual data, which we did nearly 100% automatically by speech segmentation. For each speaker of the XM2VTS database, the utterance “ 0 1 2 3 4 5 6 7 8 9” was divided into single-digit sub sequences 0 to 9. For our segmentation we used HMM models of digits Furthermore we manually verified and corrected the segmentation results so as to eliminate the impact of database segmentation errors when interpreting our recognition results. We propose two protocol setups for the XM2VTS database; protocol 1 is the well known Lausanne protocol [23], used for speaker identification and protocol 2 which is used for speech recognition. Protocol 2 is also suggested by other studies [14]. Further details of this protocol can be found in [24].

Protocol 1 – the training and the evaluation group contains 225 subjects where 200 subjects as clients and 25 subjects as impostors. For the testing group yet another 70 subjects as impostors. For clients sessions 1, 2 and 3 are used as the training and evaluation sets and session 4 as the test set. *Protocol 2* – the speakers were involved both in training SVMs and testing SVMs, we used total of 4 pronunciations for training and 4 for testing. The training and test samples were completely disjoint.

5 Classification by Support Vector Machine

The SVM formulation is based on the Structural Risk Minimization principle, which minimizes an upper bound on the generalization error, as opposed to the Empirical Risk Minimization [25][26]. An SVM is a discrimination-based binary method using a statistical algorithm. The background idea in training an SVM system is finding a hyperplane $\mathbf{w}\cdot\mathbf{x}+b=0$, as a decision boundary between two classes. For linearly separable training dataset labelled pairs $\mathbf{x}_i, y_i, i=1,\ldots,l$, where $\mathbf{x}_i \in \Re^n$ and $\mathbf{y} \in \{1,\text{-}1\}^l$, the following equation is verified for each observation data (feature vector).

$$d_i(w^T x_i + b) \geq 1 - \xi_i \ for\ i = 1, 2, ..., l\ \xi_i > 0, \tag{6}$$

where d_i is the label for sample data $\mathbf{x}_i$ which can be +1 or -1; $\mathbf{w}_i$ and b are the weights and bias that describe the hyperplane; ξ represents the number of data samples left inside the decision area, controlling the training errors. In our experiment we use the inner-product kernel function as RBF kernel

$$K(\mathbf{x},\mathbf{y}) = \exp(-\gamma \left\|\mathbf{x}-\mathbf{x}\right\|^2), \gamma > 0, \tag{7}$$

When conducting speech-classification experiments, we will need to choose between multiple classes. The best method of extending the two-class classifiers to multi-class problems appears to be application dependent. For our experiments we use the *one against one* approach. It simply constructs for each pair of classes an SVM classifier which separates those classes. All tests here were performed for only the speech signal, only the visual signal, and the merged audio-video signal by using the SVM toolkit [27].

6 Experimental results

We want to quantify the performance and audio-complementary of our visual features in speaker recognition and speech recognition using the XM2VTS database. First the text-prompted speaker-recognition test using protocol 1 are presented and then the speech-recognition system test results using protocol 2 are presented. In our experiments we use direct fusion at feature level, which are detailed in [19].

Kernel	Audio recognition rate	Visual recognition rate	Audio-Visual recognition rate
RBF	92%	80%	100%

Table 1. Speaker recognition rate by SVM using word “7”.

6.1 Speaker-identification system by SVM

A smaller dataset of 100 speakers were tested for all digits and the most significant word for the speaker recognition rate was digit “7” which gave the highest recognition rate. The experiment follows protocol 1 using all 295 speakers:

- Partition the database for training, evaluation, and testing according to protocol 1.
- Train the SVM for an utterance so that the classification score, L (the mean of the classification equation 6 for an utterance), is positive for the user and negative for impostors.
- L is compared to a threshold T.
  - Find the threshold T such that False Acceptance is equal to False Rejection using the evaluation set.
  - Using the threshold T, the decision L is made according to the rule: if $L > T$ accept the speaker else reject her/him.

However, protocol 1 is desired for verification (1:1 matching). To perform identification (1:many matching) we proceed as follows:

- Identify the speaker from a group of speakers
  - We construct classifiers to separate each speaker from all other speakers in the training set.
  - The speaker identity is determined by the classifier that yields the largest likelihood score.

Table **1** shows the results of using SVM classifiers with RBF kernel function using only one word (digit) to recognize the speaker identity. The recognition performance obtained when using coefficients both from dynamic image and speech are considerably higher than when using a single modality based on speech parameters. These results show that our features can perform well in identification problems.

6.2 Speech recognition system by SVM

In Table **2**, we illustrate all systems based on only acoustic, only visual and merged audio visual feature information. We obtain the best recognition rate for digits ”1, 6, and 7” 100%. One cause why the results in Table **2** vary is that there is not enough information (especially visual information) for certain utterances. This is not surprising because the XM2VTS database was collected for identity recognition and not speech recognition. During the segmentation we could verify

Word	Audio features	Visual features	Audio-Visual features
0	89%	70%	92%
1	90%	77%	100%
2	86%	60%	89%
3	90%	75%	96%
4	89%	55%	85%
5	90%	50%	83%
6	100%	90%	100%
7	93%	100%	100%
8	91%	54%	83%
9	90%	49%	85%

Table 2. Speech-recognition rate of all digits using protocol 2 in one against one SVM.

that when uttering the words from 0 to 9 in a sequence without silence between words, the words "4, 5, 8, 9" are pronounced in shorter time-lapses and the amount of visual data is notably less in comparison to other digits. Additionally amount of speech for each speaker differ when uttering the same word or digit depending on the manner and speed of the speaker. The average of the speech recognition over all digits is ≈ 68% and ≈ 90% for only visual and only audio system respectively.

7 Conclusion and discussion

In this paper we described a system utilizing lip movement information in dynamic image sequences of numerous speakers for robust speech and speaker recognition by no use of iterative algorithm or assuming successful lip-contour tracking. In environments such as airports, outside traffic, train station etc. the automatic speech recognition or speaker recognition system based on only acoustic information would with high probability be unsuccessful. Our experimental results support the importance of adding lip motion representation in speaker- or speech-recognition systems that can be installed for instance in mobile devices as a complement to acoustic information.

We presented a novel lip-motion quantization and recognition results of lip-motion features as standalone and as a complement to audio for speaker and speech recognition tasks using extensive tests. Significant improvements of audio based recogntion utilizing our motion features to achieve high recognition performance, for speech as well as identity are provided.

We noted via our segmentation that the words "4, 5, 8, 9" were containing less visual information during the speech utterance in the XM2VTS database. The poor recognition performance of these digits indicate that XM2VTS database does not contain sufficient amounts of visual information on lip movements. Not surprisingly, if the visual feature-extraction is made on sufficient amount of visual speech data, the available modelling for recognition tasks appears to be sufficient for successful recognition.

References

1. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.: Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE **91**(9) (2003) 1306–1326
2. Brunelli, K.R., Falavigna, D.: Person identification using multiple cues. IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(10) (1995) 955–966
3. Chibelushi, C., Deravi, F., Mason, J.: A review of speech-based bimodal recognition. IEEE Transactions on Multimedia **4**(1) (2002) 23–37
4. Duc, B., Fischer, S., Bigun, J.: Face authentication with sparse grid gabor information. IEEE International Conference Acoustics, Speech, and Signal Processing **4**(21) (1997) 3053–3056
5. Tang, X., Li, X.: Video based face recognition using multiple classifiers. Sixth IEEE International Conference on Automatic Face and Gesture Recognition $FGR2004$ - IEEE Computer Society (2004) 345–349
6. Faraj, M.I., Bigun, J.: Person verification by lip-motion. 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) (2006) 37–45
7. Luettin, J., Maitre, G.: Evaluation protocol for the extended m2vts database *xm2vtsdb*. (1998) in: IDIAP Communication 98-054, Technical report R R-21, number = IDIAP - 1998.
8. Dieckmann, U., Plankensteiner, P., Wagner, T.: Acoustic-labial speaker verification. Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, LNCS 1206 (1997) 301–310
9. Jourlin, P., Luettin, J., Genoud, D., Wassner, H.: Acoustic-labial speaker verification. Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, LNCS 1206 (1997) 319–326
10. Chen, T.: Audiovisual speech processing. IEEE Signal Processing Magazine **18**(1) (2001) 9–21
11. Liang, L., Zhao, X.L.Y., Pi, X., Nefian, A.: Speaker independent audio-visual continuous speech recognition. IEEE International Conference on Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 **2** (2002) 26–29
12. Kollreider, K., Fronthaler, H., Bigun, J.: Evaluating liveness by face images and the structure tensor. In AutoID 2005: Fourth Workshop on Automatic Identification Advanced Technologies - IEEE Computer Society (2005) 75–80
13. Wan, V., Campbell, W.: Support vector machines for speaker verification and identification. Proceedings of the 2000 IEEE Signal Processing Society Workshop, Neural Networks for Signal Processing X. **2** (2000) 775–784
14. Gavat, I., Costache, G., Iancu, C.: Robust speech recognizer using multiclass svm. 7th Seminar on Neural Network Applications in Electrical Engineering. NEUREL 2004 (2004) 63–66
15. Clarkson, P., Moreno, P.: On the use of support vector machines for phonetic classification. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP. **2** (1999) 585–588
16. Reynolds, D., Quatieri, T., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing **10**(1–3) (2000) 19–41
17. Farrell, K., Mammone, R., Assaleh, K.: Speaker recognition using neural networks and conventional classifiers. IEEE-Computer Society Press **2**(1) (1994) 194–205
18. Bigun, J., Granlund, G., Wiklund, J.: Multidimensional orientation estimation with applications to texture analysis of optical flow. IEEE-Trans Pattern Analysis and Machine Intelligence **13**(8) (1991) 775–790

19. Faraj, M.I., Bigun, J.: Audio-visual person authentication using lip-motion from orientation maps. Submitted to publication (2006)
20. Granlund, G.H.: In search of a general picture processing operator. Computer Graphics and Image Processing **8**(2) (1978) 155–173
21. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on Acoustics, Speech, and Signal Processing **28**(4) (1980) 357–366
22. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The htk book (for htk version 3.0). (2000) http://htk.eng.cam.ac.uk/docs/docs.shtml.
23. Messer, K., Matas, J., Kittler, J., Luettin, J.: Xm2vtsdb: The extended m2vts database. In Second International Conference of Audio and Video-based Biometric Person Authentication, $ICSLP'96$ (1999) 72–77
24. Faraj, M.I., Bigun, J.: Visual motion representation and acoustic constraints for bi-modal biometric speech recognition and speaker verification. Invited paper, submitted to publication (2007)
25. Vapnik, V.N.: The Nature of Statistical Learning Theory. (1995) Springer.
26. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery **2**(2) (1998) 121–167
27. Chang, C.C., Lin, C.J.: Libsvm–a library for support vector machines. software available at www.csie.ntu.edu.tw/ cjlin/libsvm (2001)