Adapting a High Quality Audiovisual Database to PDA Quality

Kevin McTait, Hervé Bredin, Silvia Colón, Thomas Fillon and Gérard Chollet
GET-ENST CNRS-LTCI
46 rue Barrault
75634 Paris cedex 13, France
{*mc-tait,bredin,colon,fillon,chollet*}*@tsi.enst.fr*

Abstract

Audiovisual sequences in the BANCA database are degraded in quality so that they resemble audiovisual sequences recorded using the in-built sensors of a handheld PDA device. The quality of the new database is therefore reduced on a number of levels, but the recordings remain equivalent. This process is applicable for any given high quality audiovisual database and set of lower quality sensors. More specifically, this work has been carried out in the context of the SecurePhone project, the aim of which is to develop a multimodal biometric identity verification system embedded on a PDA device. Initial results show that this adaptation process produces adequate audiovisual sequences for a given PDA device, removing the need to record a training database for each new PDA device.

1. Introduction

The goal of the SecurePhone project is the realisation of a multimodal biometrics-based identity verification system to be embedded in a mobile device enabling (biometrically) authenticated users to deal legally binding m-contracts during a mobile phone call in an easy yet highly dependable and secure way. In the case of the SecurePhone project, the PDA/mobile device chosen is the Qtek 2020 smartphone.

A database resembling the operating conditions of the smartphone, given all biometric modalities, is vital to the success of the SecurePhone project. Such databases are necessary for:

- Development and testing of competing biometric authentication algorithms.
- Establishing formal evaluation protocols and benchmarks given the final prototype.

In the context of the SecurePhone project, benchmark databases, covering each of the four modalities of the proposed biometric recogniser, are to be adapted to resemble the operating conditions of the smartphone. In addition, a novel PDA database is to be recorded in order to complement the adapted databases and test the performance of the database adaptation process. The four modalities addressed in the project are:

- Face verification using one or more image frames from a video recording of the subject
- Speaker verification using audio signals
- Online handwritten signatures verification
- Speaking faces identity verification using synchronised audiovisual sequences

Adapting audiovisual sequences for the latter modality is the subject of this article. The benchmark audiovisual database chosen for the project is the BANCA database [3]. The resulting software or adaptation script is to be made available as open source software, particularly since many of the function calls are made to existing open source software packages, libraries or toolkits.

The remainder of this article is organised as follows: Section 2 outlines the motivations and possible weaknesses of adapting an audiovisual database for biometric identity verification, Section 3 outlines the adaptation methodology, Section 4 a prospective evaluation methodology and finally our conclusions are detailed in Section 5.

2. Motivation

In order to have a database resembling the operating conditions of the Qtek smartphone, an obvious solution would be to record a suitable database using the sensors of the Qtek device. However, this is undesirable for several reasons:

- *Benchmarking*: Results obtained on an adapted benchmark database available to the research community, as is the case for BANCA, enables comparison of results among state of the art systems (particularly since the adaptation script is to be made available as open source software).
- *Lifespan*: As the quality of the sensors embedded in mobile devices improve rapidly over time, a database recorded using the Qtek smartphone would have a limited lifespan. This database would then only serve the

requirements of the SecurePhone project and would soon become obsolete. In the same way, one may choose to replace the camera or microphone, resulting in a different signal quality. Adaptation therefore allows the original BANCA database to be adapted to any other (future) PDA or smartphone device (as this article attempts to show).

- *Synchronisation*: Without low-level APIs, the Qtek 2020 does not allow for synchronised audiovisual recordings, necessary for 'speaking faces' identity verification and robustness against impostor attacks. BANCA and other high quality audiovisual databases do provide the required level of synchrony.

Adaptation is a proven process for audio signals. In the context of speaker verification, a typical forgery scenario includes automatic voice transformation techniques that an impostor may use to assume the identity of an authorised client. For GMM-based speaker verification it may be sufficient to find a mapping function F between the impostor's feature vectors x and those of the client y [2] [7] [8] [5]. Given two sequences composed by the same words, pronounced respectively by the impostor and by the client, F can be derived by minimizing the mean square error where E is the expectation (1). This approach may require only a relatively small amount of client data, but it has the disadvantage of being text, language and speaker dependent.

$$\epsilon_{mse} = E[||y - F(x)||^2] \tag{1}$$

A more effective and language, text and speaker independent approach to voice conversion is based on the ALISP (Automatic Language Independent Speech Processing) approach [6]. In this approach, a database of a client's speech segments (arbitrary segments or ALISP units acquired in an unsupervised manner that approximate manually defined phones) is maintained. Subsequently, recognition of the impostor's speech in terms of ALISP units allows the replacement of the impostor's voice segments with equivalent representative units taken from the client's codebook. The HMM-based recognition performed in the ALISP processing system assures that a subsequent recognition performed by the verification system will result in a good phonetic match. Recent experiments [11] have shown that using ALISP as a voice forgery technique returns significant results.

In the same way that acoustic units of an impostor's voice are replaced by acoustic units of a client's voice using ALISP, video segments are equally adaptable and replaceable. The impostor maintains a database of face and speech feature vectors of a client which are then used to drive and animate an MPEG-4 compliant face model [10], [9]. Adaptation of video sequences is not only restricted to forgery scenarios but also to recognition and synthesis of facial expressions [1] and even speech-driven facial animation [4] [13].

However, in terms of audiovisual biometric identity authentication, it remains to be seen how well the adaptation process works. For example, there is a risk that a biometrics based verification system trained and developed on an adapted database would result in a discrepancy between this training data and any test data acquired directly on the smartphone in a real world test scenario since machine recognition is extremely sensitive to the channels used for data recording. The ideal scenario, would be to compare results obtained using the adapted BANCA database with an equivalent database recored directly using the Qtek 2020 sensors. If there is no major discrepancy between the results, then the adaptation process may be considered a success. Evaluation is considered in more detail in Section 4.

3. Adaptation Methodology

3.1. PDA Specifications

In order to adapt an existing audiovisual database to resemble a database recorded using a specific device, it is intuitive that the specifications of the device in question be known. In terms of the Qtek 2020, the maximum video size available is 240 x 320 pixels and the resulting video sequences are available in either MPEG-4, Motion-JPEG AVI or H.263 video conferencing format. The frame rate of the Qtek smartphone is 10 frames per second.

In terms of audio capture, the audio signals recorded simultaneously with the video data are stored as PCM files, sampled at 8 kHz, single channel with 16 bit precision. Without low-level APIs, it would appear that it is not possible to change the specifications of the audio channel and select one of the possible higher sampling rates (up to 44kHz).

3.2. Adaptation Tools

In order to distribute the adapted BANCA database, it is envisaged that the adaptation script used will be distributed as open source software. Given this situation, we have ensured that the adaptation procedure largely makes use of open source software making the script portable. In the case of user defined functions, these will also be bundled with the adaptation script. Distributing adaptation software, as opposed to an adapted database is preferable for several reasons:

- The logistical problem of distributing a database of potentially many gigabytes of data is removed.
- The legal problems involved in distributing a proprietary database are removed.
- Given the size and complexity of database adaptation, the process is considered incremental. Algorithms and techniques may be improved over time to tackle the more challenging problems involved. More importantly, the parameters of the script may be modified to take into account adaptation from different or new high quality databases to different PDA sensors or devices.

There is a large amount of open source audio and video file processing software, most of which concerns the conversion from and to various video formats. For the most part, the adaptation methodology presented in this article makes use of open source toolkits or audiovisual software for the reasons outlined above.

The *transcode* [1] software package was largely used to process the video files with image manipulation handled by user defined C functions within the *Open Computer Vision Library* [2] which is a collection of algorithms and sample code for various computer vision problems. Audio processing is handled either by *transcode*, where appropriate, or by *Sox* [3].

The final adaptation script is of the form of a perl script running under either Unix or Linux that takes a list of AVI files as input and calls the various functions from the open source software or other user defined code, as described above. The output is a list of adapted AVI files.

3.3. Methodology

In order to obtain audiovisual sequences from the BANCA database, GET-ENST acquired copies of the original DV tapes used to recorded the BANCA subjects (the published database contains the full audio recordings but only a given number of still image video frames per client access). The resulting full video BANCA sequences are available at 25 frames per second, the video size is 720 x 576 pixels and the audio sampling rate 48 kHz (PCM, 16 bit precision).

In order to adapt the BANCA audiovisual sequences to the specifications of the PDA sensors, several subtasks were identified (some remain prospective or under study). The following sections outline these subtasks:

3.3.1 Frame rate reduction

The first stage in the adaptation process is to reduce the number of frames per second in the original BANCA sequence to match that of the PDA. The original BANCA AVI sequences are therefore converted from 25 frames per second to the maximum 10 frames per second possible by the Qtek PDA. This represents a simple operation with *transcode*:

```
transcode -i infile.avi -o outfile.avi
--export_fps 10 -Jmodfps -k -z
-xffmpeg -yffmpeg -Fmpeg4
```

3.3.2 Size reduction

Subsequently, this new AVI file is reduced in size and scaled just larger than the maximum PDA video size of 240x320 pixels. Scaling the image size larger than the maximum video size of the PDA facilitates the production of the effect of camera shake (see Section 3.3.4). The smallest possible scaling factor is desired since reducing the resolution introduces a process where the colours of the original pixel area are averaged to form the reduced pixel area, potentially introducing new colours. Size reduction is performed once again with *transcode*:

```
transcode -i infile.avi -o outfile.avi
-k -z -xffmpeg -yffmpeg -Z480x384 -Fmpeg4
```

3.3.3 Face localisation

Subsequent steps of the adaptation process (camera shake, colour reduction) require the conversion of the AVI file into its constituent still image frames. Each subsequent frame is then processed. *transcode* is again used to break down the AVI into video frames:

```
transcode -i infile.avi -o out/frame1
-k -z -x ffmpeg -y ppm
```

In this case, the still image frames are converted into PPM format, (starting frame0001.ppm ...) and stored in the directory named *out*.

In order to simulate PDA usage conditions in the context of the SecurePhone project where the face of the client largely fills the screen in a central position (as shown in Figure 1), it is necessary to use an automatic face localisation algorithm [14]. A front-face localisation algorithm based on [12] available in *OpenCV* has been used. The coordinates of the detected face is therefore used to centre the subject in the image frame.

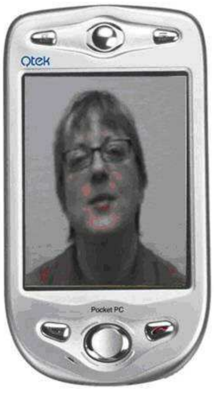

Figure 1. Speaker on the PDA screen

3.3.4 Camera shake

A C program was written that defines a new 240x320 bounding box with the eyes as the centre. The fact that this bounding box is smaller than the scaled bounding box enables us to simulate the effects of camera shake in a hand held portable device.

[1] http://zebra.fh-weingarten.de/∼ transcode/

[2] http://sourceforge.net/projects/opencvlibrary

[3] http://sox.sourceforge.net

In order to simulate camera shake, the x and y coordinates of the 240x320 bounding box are varied in n predefined directions, around a central point within a certain variance. The direction of movement is controlled by a random variable subsequently returning to the central point, according to the natural human reflex to recentre one's image, before a different movement direction is defined by the random variable.

Currently, the variance of the x and y coordinates is defined manually and remains constant. However, we have already begun experiments to automatically model the variance given the motion of the detected faces (with *OpenCV* face localisation algorithm) in training sequences of subjects using the PDA as envisaged in the SecurePhone project.

3.3.5 Colour reduction

The size of the colour space given the PDA camera is reduced compared to that available in the original BANCA sequences. The range of colours available in the BANCA video sequences must therefore be reduced in order to resemble video sequences recorded using the inbuilt PDA camera.

In order to define a colour mapping function between images taken using two different cameras, equivalent (if not identical) training images are required, recorded using both cameras. Since we cannot replicate BANCA using the PDA camera (no access to the original subjects nor recording situation), it was decided to record equivalent training images using the PDA camera and a high quality video camera similar to that used to record the original BANCA database providing video output with comparable if not identical specifications to that of the original BANCA video sequences.

Training frames of TV style test patterns, as shown in Figure 2, were recorded with both the PDA camera and a high quality commercial video camera. This enabled a broad range of colours to be present with the aim of establishing a correspondence between the colour ranges of the two cameras.

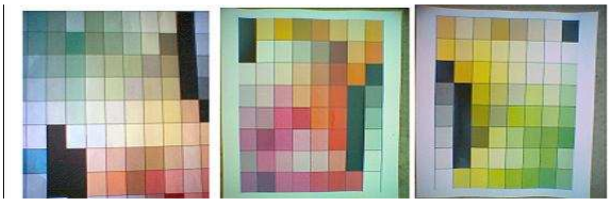

Figure 2. TV style testcards

In order to define the mapping function between the two colour spaces, the values of each equivalent colour zone (each square) from the low and high quality recordings of the test patterns were extracted and mapped onto a two dimensional space: for each given colour (square) the colour values of the low quality frames were plotted on the x axis and the values of the high quality frames on the y axis. Three colour spaces, RGB, HSV and YCrCb, were studied in order to best define a clear correspondence between the colours of the training frames. In each case, plots were established for each component of a given colour space. Figure 3 shows the graphs obtained.

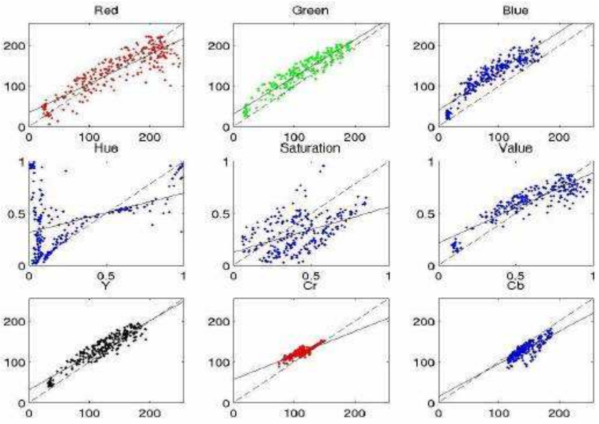

Figure 3. Colour mapping functions for 3 colour spaces

In Figure 3, the dotted lines indicate the function $y = x$ or a direct correspondence between the colour values between high and low quality training frames. The solid line represents the function of best fit or the best possible correspondence between colour values of the training frames. This function is polynomial and of the form $y = ax + b$ and was calculated using a standard interpolation function, in this case, the predefined polynomial function *Polyfit* from the Matlab programming language.

Figure 3 also shows that the most convenient colour space available from which the mapping function may be defined is RGB (Red, Green, Blue) since more colour values are present along the totality of the x and y axes and they represent a largely linear (hence easier to define) relationship.

However, this polynomial function is presumed to be sensitive to illumination and other factors affecting recording conditions. Therefore, the parameters may have to be modified according to a given recording scenario.

After the colour space of each video frame is modified, the PPM video frames are recombined into a new AVI file of size 240 x 320 pixels at 10 frames per second. Again, *transcode* was used for this purpose:

```
transcode -i ppm_in/list.txt -ximlist, null
-g240x320 -yffmpeg,null -f10 -k -z -H0
-o avi_out -Fmpeg4
```

where the input to the command is a sorted list of the PPM files and the new AVI file is output to a directory named *avi_out* in this case.

3.3.6 Audio downsampling

The audio stream was extracted from the AVI file and downsampled from 44 kHz to 8 kHz. While Sox is frequently used for this purpose, *transcode* is equally capable and was preferred in this case to reduce the number of software packages required for the adaptation process:

```
transcode -i avi_in -o audio_out -Jresample
-E8000 -e48000 -x ffmpeg -y null,wav -N 0x1
```

where the output is a specified as a wave file.

Finally, the downsampled audio stream and the adapted video stream are remerged to form the final adapted audiovisual sequence:

```
transcode -i video_in -o avi_out -p audio_in
-k -z -x ffmpeg -y ffmpeg -F mpeg4
```

3.4. Prospective Studies

Given that the task of adapting audiovisual signals to match the specifications of inferior sensors as closely as possible is a sizeable task, many subtasks have presently remained outside the scope of this paper. However, these prospective but important areas of study are currently under investigation:

3.4.1 Optical distortion

Due to the difference between the lens of a high quality video camera and that of the PDA camera, an amount of optical distortion is inevitable. Compared to the training frames from BANCA, the images taken by the PDA camera display an amount of radial distortion, particularly with recordings taken at close range, as shown in Figure 4. To deal with this problem, *OpenCV* comes with a C implementation of *Camera Calibration Toolbox for Matlab* [4] allowing the computation of distortion parameters for any camera. Optical transformation is then computed in two steps. The high quality video camera image is undistorted to give a temporary image, which is then distorted according to the learned parameters of distortion of the PDA camera.

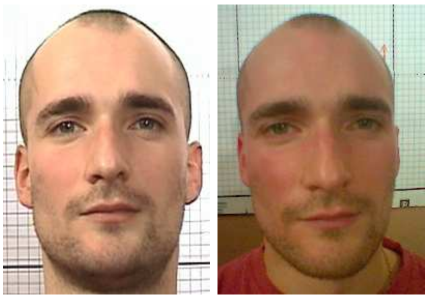

Figure 4. Optical distorsion
Left: high quality video camera. Right: PDA camera.

3.4.2 Sensor noise

In order to introduce the correct amount of noise into the adapted video sequences, the amount of noise must first be calculated. While theoretically straightforward, this represents a fairly complicated procedure practically in that absolutely uniform images (in terms of colour) must be taken by the low quality camera to produce flat histograms where each pixel has the same value, from which the peak values (interpreted as noise) may be calculated. Practically, this is extremely difficult to produce.

In a perfectly uniform image, where all pixels have the same value, the standard deviation from the average of all pixels (2) is null. However, in a practical situation, this formula will not be able to distinguish between noise pixels and non-uniform zones in the image.

$$\text{standard deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{2}$$

3.4.3 Noisy backgrounds

Given the limited background conditions of the BANCA database, it may be possible to introduce noisy backgrounds or backgrounds of choice (such as outdoor scenes etc) to simulate mobile conditions as required by the SecurePhone project. As this requires the use of face localisation techniques, the result of which may not always be 100% accurate, this subtask remains the subject of prospective work.

3.4.4 Microphone discrepancies

Given that any mismatch between training and test data (in terms of acoustic conditions, channel, speaker gender, sampling frequency, microphone type etc.) is known to strongly affect the error rates in speaker verification, database adaptation should also take into account any discrepancies in the frequency response between the two microphones used to record the original BANCA database (the high and low quality microphone) and the PDA microphone.

In the same way, any differences between the original microphones' pick-up pattern and that of the PDA microphone will affect the amount of room echo and background noise that is picked up. The effect of both these discrepancies should be evaluated and introduced into the adaptation procedure to minimise any mismatch between training and test data.

4. Evaluation and Results

There is no well known procedure for evaluating the proximity of an adapted database to that of equivalent recordings taken by the PDA in question. However, in the case of the SecurePhone project (biometrics based identity verification), one could envisage a prospective evaluation methodology given that, in the context of the project, the recording of a new database recorded using the in-built sensors of the Qtek 2020 is currently underway, but not yet completed:

World models are trained using the adapted audiovisual sequences from BANCA. Subsequently, client models are trained and tested using the data from the database

[4] http://www.vision.caltech.edu/bouguetj/calib_doc/

recorded using the PDA sensors. Conversely, world models are trained on the data from the PDA database, while client models are trained and tested on the data from the adapted BANCA database. The results obtained from the two steps may be compared. If there are no major discrepancies between the results of the two tests, then adaptation may be considered as a success. In the same manner, any future improvements to the adaptation process (the fact that the process may be improved continually over time) will be evaluated using this method. As the discrepancy between the results of the two evaluations closes, the adaptation process may be considered more successful.

A further and similar evaluation would be to record a small test database, of a limited number of subjects for testing purposes only, simultaneously using a high quality video camera and the PDA sensors. In this way equivalent (near identical) high and low quality audiovisual sequences are obtained. The high quality sequences are degraded using the procedure outlined in this article. Audiovisual biometric identity verification tests are subsequently carried out, first on the PDA sequences and then on the adapted high quality sequences. The difference between the results demonstrates the success of the adaptation procedure. Again, as the adaptation process improves, so should the difference between the two results should close.

5. Conclusions

We have shown that it is possible to adapt or rather degrade high quality audiovisual signals to represent audiovisual sequences recorded using a handheld portable device with inferior capture devices (camera, microphone).

We have also shown how this procedure is applicable to all types of device, given the specifications of the two capture devices / audiovisual sequences and a small amount of training data.

In this way, the recording of new audiovisual training database for every new PDA device on the market is no longer necessary, reducing the development time for embedded PDA applications such as the multimodal biometric identity authentication system specified by the SecurePhone project.

However, the procedure used to adapt the colour space (Section 3.3.5) remains extremely sensitive to changes in illumination, but given that this remains stable for each of the three recording conditions in BANCA (controlled, degraded, adverse) this is not a problem for database adaptation in the SecurePhone project. Initial studies into the process of adaptation of optical distortion between two lenses of differing qualities also reveals that the level of distortion is highly sensitive to the distance between subject and camera.

6. Acknowledgments

This work has been funded by the European *SecurePhone* project no. IST-2002-506883. Thanks also to Guido Aversano and Bouchra Abboud for their contributions to the work reported in this article.

References

[1] B. Abboud, F. Davoine, and M. Dang. Facial Expression Recognition and Synthesis based on an Appearance Model. *Signal Processing: Image Communication*, 10(8):723 – 740, 2004.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice Conversion Through Vector Quantization. In *Proceedings ICASSP 88*, pages 655 – 658, New York, 1988.

[3] E. B. Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J. P. Thiran. The BANCA Database and Evaluation Protocol. In *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, pages 625 – 638, Guildford, UK, 2003.

[4] C. Bregler, M. Covell, and M. Slaney. Video Rewrite: Driving Visual Speech with Audio. In *Proceedings ACM SIGGRAPH 97*, pages 353 – 360, 1997.

[5] O. Cappe, Y. Stylianou, and E. Moulines. Statistical Methods for Voice Quality Transformation. In *Proceedings of EUROSPEECH 95*, pages 447 – 450, Madrid, Spain, 1995.

[6] G. Chollet, J. Cernocky, A. Constantinescu, S. Deligne, and F. Bimbot. Toward ALISP: Automatic Language Independent Speech Processing. In K. Ponting and R. Moore, editors, *Computational Models for Speech Pattern Processing*, pages 375 – 387. Springer Verlag, 1999.

[7] A. Kain and M. Macon. Spectral Voice Conversion for Text to Speech Synthesis. In *Proceedings ICASSP 98*, pages 285 – 288, Seattle, WA, 1998.

[8] A. Kain and M. Macon. Design and Evaluation of a Voice Conversion Algorithm Based on Spectral Envelope Mapping and Residual Prediction. In *Proceedings of ICASSP-2001*, Salt Lake City, USA, 2001.

[9] W. Karam, C. Mokbel, G. Aversano, C. Pelachaud, and G. Chollet. An Audiovisual Imposture Scenario by Talking Face Animation. In G. Chollet, A. Esposito, M. Faundez, and M. Marinaro, editors, *to appear in Nonlinear Speech Processing: Algorithms and Analysis*. Springer-Verlag, 2005.

[10] S. Pasquariello and C. Pelachaud. Greta: A Simple Facial Animation Engine. In *6th Online World Conference on Soft Computing in Industrial Applications, Session on Soft Computing for Intelligent 3D Agents*, 2001.

[11] P. Perrot, G. Aversano, G. Chollet, and M. Charbit. Voice Forgery Using ALISP: Indexation in a Client Memory. In *Proceedings of ICASSP 2005*, Philadelphia, PA, 2005.

[12] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. *IEEE Computer Vision and Pattern Recognition*, 1:511 – 518, 2001.

[13] J. Williams and A. Katsaggelos. An HMM-Based Speech-to-Video Synthesizer. *IEEE Transactions on Neural Networks*, 13(4):900 – 915, 2002.

[14] M. Yang, D. Kriegman, and N. Ahuja. Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34 – 58, 2002.