IDIAP COMMUNICATION

THE VIDTIMIT DATABASE

Conrad Sanderson (*)

IDIAP–Com 02-06

AUGUST 2002

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

(*) conradsand @ ieee.org

1 Abstract

This communication describes the multi-modal VidTIMIT database, which can be useful for research involving mono- or multi-modal speech recognition or person authentication. It is comprised of video and corresponding audio recordings of 43 volunteers, reciting short sentences selected from the NTIMIT corpus [8].

2 Introduction

At the start of research for my thesis [14] on multi-modal person authentication, only one widely distributed multi-modal database existed, namely the M2VTS database [10]. The database is comprised of video sequences and corresponding audio recordings of 37 people counting '0' to '9' in their native language (mostly in French). There are five sessions per person (with one '0' to '9' utterance per session), spaced apart by at least one week. A head rotation sequence was also recorded during each session, where each person moved their head to the left and then to the right. The head rotation is meant to facilitate extraction of profile or 3D information.

The major drawbacks of the M2VTS database are its small size and the very limited vocabulary (one "phrase" consisting of the '0' to '9' count). The small size results in several problems. The data set needs to be divided into at least 2 sections, representing the *training* and *testing* sections (typically, M2VTS sessions 1 to 3 are labeled as training data and session 4 as test data, with session 5 left out due to particular recording conditions). A small amount of training data can easily result in unreliable statistical models [5]. A small test set results in a small number of verification tests, thus any relative improvement of one verification approach over another is dubious. Lastly, a verification method developed on the M2VTS database cannot be guaranteed to work in the more general text-independent mode, since the training phrase is the same as the testing phrase.

The Extended M2TVS (XM2VTS) database [9], released several years later, addresses some of these problems. The main differences are: 295 subjects, three fixed phrases (with two utterances of each phrase) and four sessions. The phrases are:

1. "0 1 2 3 4 5 6 7 8 9"
2. "5 0 6 9 2 8 1 3 7 4"
3. "Joe took fathers green shoe bench out"

While the number of subjects results in a much larger number of verification tests, the database is inherently suited for development of text-dependent verification systems. While it is possible to obtain a pseudo text-independent setup by training a system using only phrases 1 and 2 and testing it on phrase 3, the training data is hardly representative of the test data - easily leading to poor performance.

At the time of release, the XM2VTS database was quite expensive to obtain; moreover, it was distributed on DVD-RAM media at a time when the DVD-RAM drives were quite expensive and not widely available. Due to financial limitations, I was not able to obtain the XM2VTS database.

Taking into account the problems with the M2VTS and XM2VTS databases, I have created the VidTIMIT database[1], described in the following section.

[1]The VidTIMIT database was created while I was a PhD student at Griffith University, Australia, under the supervision of Professor Kuldip K. Paliwal.

3 VidTIMIT Database

The VidTIMIT database is comprised of video and corresponding audio recordings of 43 volunteers (19 female and 24 male), reciting short sentences. It was recorded in 3 sessions, with a mean delay of 7 days between Session 1 and 2, and 6 days between Session 2 and 3.

The delay between sessions allows for changes in the voice, hair style, make-up, clothing and mood (which can affect the pronunciation), thus incorporating attributes which would be present during the deployment of a verification system. Additionally, the zoom factor of the camera was randomly perturbed after each recording.

The sentences were chosen from the test section of the NTIMIT corpus [8]. There are 10 sentences per person. The first six sentences (sorted alpha-numerically by filename) are assigned to Session 1. The next two sentences are assigned to Session 2 with the remaining two to Session 3. The first two sentences for all persons are the same, with the remaining eight generally different for each person. The mean duration of each sentence is 4.25 seconds, or approximately 106 video frames.

A typical example[2] of the sentences used is in Table 1. There is complete correspondence of the subject IDs between VidTIMIT and NTIMIT (and hence the recited sentences).

In addition to the sentences, each person performed an extended *head rotation* sequence in each session, which allows for extraction of profile and 3D information. The sequence consists of the person moving their head to the left, right, back to the center, up, then down and finally return to center.

The recording was done in a noisy office environment (mostly computer fan noise) using a broadcast quality PAL digital video camera. The video of each person is stored as a numbered sequence of JPEG images with a resolution of 384 × 512 pixels (rows × columns). 90% quality setting was used during the creation of the JPEG images. The corresponding audio is stored as a mono, 16 bit, 32 kHz WAV file. The entire database occupies approximately 3.5 Gb. Online video examples are available at: *http://www.idiap.ch/~sanders/vidtimit/*

It must be noted that unlike the M2VTS and XM2VTS databases, all sessions contain various phonetically balanced sentences. If we define Session 1 as the training section and Sessions 2 & 3 as the test section then no sentences are repeated across the test and train sections; the database is thus suited for the development of a text-independent verification system.

Examples images of two subjects are shown in Figure 1. The first, second and third columns represent images taken in Session 1, 2 and 3, respectively.

4 Database Structure

Let us assume that the database is stored in the *vidtimit* directory. The video for each subject is stored as:

vidtimit/video/subjectID/sentenceID/###

where *subjectID* is the subject ID (e.g., *felc0*), *sentenceID* is the head rotation or sentence ID (e.g., *sx396*) and *###* is a three digit frame ID (e.g., *037*). Each frame is stored as a JPEG file. The corresponding audio[3] is stored as a WAV file in:

vidtimit/audio/subjectID/sentenceID.wav

[2]Copyright restrictions on the NTIMIT corpus prevent the list of all sentences used in the VidTIMIT database.

[3]It must be noted that there is no audio for the head rotation sequences.

Section ID	Sentence ID	Sentence text
Session 1	sa1	She had your dark suit in greasy wash water all year
	sa2	Don't ask me to carry an oily rag like that
	si1398	Do they make class-biased decisions?
	si2028	He took his mask from his forehead and threw it, unexpectedly, across the deck
	si768	Make lid for sugar bowl the same as jar lids, omitting design disk
	sx138	The clumsy customer spilled some expensive perfume
Session 2	sx228	The viewpoint overlooked the ocean
	sx318	Please dig my potatoes up before frost
Session 3	sx408	I'd ride the subway, but I haven't enough change
	sx48	Grandmother outgrew her upbringing in petticoats

Table 1: Typical example of sentences used in the VidTIMIT database.

Figure 1: Example of subjects in the VidTIMIT database. The first, second and third columns represent images taken in Session 1, 2 and 3, respectively.

5 Proposed Protocols for Person Authentication/Verification Experiments

5.1 Background

Since the verification system is inherently a two-class decision task, it follows that the system can make two types of errors. The first type of error is a False Acceptance (FA), where an impostor is accepted. The second error is a False Rejection (FR), where a true claimant is rejected. Thus the performance is measured in terms of False Acceptance rate (FA%) and False Rejection rate (FR%), defined as:

$$\text{FA\%} = \frac{I_A}{I_T} \times 100\% \quad (1)$$

$$\text{FR\%} = \frac{C_R}{C_T} \times 100\% \quad (2)$$

where I_A is the number of impostors classified as true claimants, I_T is the total number of impostor classification tests, C_R is the number of true claimants classified as impostors, and C_T is the total number of true claimant classification tests.

Since the errors are related, minimizing the FA% increases the FR% (and vice versa). The trade-off between FA% and FR% can be adjusted using a threshold (see [1] or [11] for an example). Depending on the application, more emphasis may be placed on one error over the other. For example, in a high security environment, it may be desired to have the FA% as low as possible, even at the expense of a high FR%.

There seems to be two schools of thought for measuring the performance of a verification system. In the first method, the trade-off between FA% and FR% can be graphically represented by a Receiver Operating Characteristics (ROC) plot or a Detection Error Trade-off (DET) plot [1, 4] (where the results can be interpreted more easily than on the ROC plot). In both cases the FR% is plotted as a function of FA%. To quantify the performance into a single number, Equal Error Rate (EER), is often used [7]; here the system is configured to operate with FA% = FR%.

It must be noted that in this method, the decision threshold (or equivalent decision mechanism) is adjusted to obtain desired performance on *test* data[4] (data unseen by the system up to this point). Such a threshold is known as the *a posteriori* threshold.

In the second method, the decision threshold is fixed before finding the performance; it is known as the *a priori* threshold [6]. The *a priori* threshold can be found via experimental means using training data or *evaluation* data (data which has also been unseen by the system up to this point, but is separate from *test* data).

Logically, the *a priori* threshold is more realistic. However, it is often difficult to find a reliable *a priori* threshold [4, 6]. In the second method, the database is often divided into three sets: *training* data, *evaluation* data and *test* data. If the *evaluation* data is not representative of the *test* data, then the *a priori* threshold will achieve significantly different results on *evaluation* and *test data*. Moreover, compared to using the first method, such a database division reduces the number of verification tests, thus decreasing the statistical significance of the results. For these reasons, many researchers use the *a posteriori* threshold and interpret the performance obtained as the *expected* performance.

[4] In the first method, the database used for experiments is usually divided into two sections: *training* data and *test* data.

5.2 Proposed Protocol I: A Posteriori Performance

In this protocol, Session 1 of the database is used for training the client models, while Sessions 2 and 3 are treated as the *test section* and are used to find the verification performance using the *a posteriori* decision threshold (or equivalent decision mechanism). Moreover, subjects *fadg0*, *faks0*, *fcft0*, *fcmh0*, *mstk0*, *mtas1*, *mtmr0* and *mwbt0* (i.e., 4 female and 4 male) are to be used only for impostor tests; this leaves 35 subjects for true claimant tests. In total, there are 1120 ($35 \times 4 \times 8$) impostor and 140 (35×4) true claimant tests. Publications using this protocol: [12, 13, 14].

5.3 Proposed Protocol II: A Priori Performance Type A

In this protocol, Session 1 of the database is used for training the client models, Session 2 is used to find the *a priori* decision threshold (or parameters for an equivalent decision mechanism) and Session 3 is used to find the final performance (for example, in terms of Half Total Error Rate [1, 3]). The meaning of Sessions 2 and 3 is then swapped, i.e., Session 3 is used to find the *a priori* decision threshold and Session 2 is used to find the performance. The two performance figures are then averaged. As for Protocol I (described above), subjects *fadg0*, *faks0*, *fcft0*, *fcmh0*, *mstk0*, *mtas1*, *mtmr0* and *mwbt0* (i.e., 4 female and 4 male) are to be used only for impostor tests. Thus in total there 1120 (560×2) impostor and 140 (70×2) true claimant tests.

5.4 Proposed Protocol III: A Priori Performance Type B

In this protocol, the database is first divided into two sections: the development section and the evaluation section. Subjects *fadg0*, *faks0*, *fcft0*, *fcmh0*, *fcmr0*, *fcrh0*, *fdac1*, *fdms0*, *fdrd1*, *mabw0*, *mbdg0*, *mbjk0*, *mccs0*, *mcem0*, *mdab0*, *mdbb0*, *mdld0*, *mgwt0*, *mjar0*, *mjsw0* and *mmdb1* are assigned to the development section, while subjects *fedw0*, *felc0*, *fgjd0*, *fjas0*, *fjem0*, *fjre0*, *fjwb0*, *fkms0*, *fpkt0*, *fram1*, *mmdm2*, *mpdf0*, *mpgl0*, *mrcz0*, *mreb0*, *mrgg0*, *mrjo0*, *msjs1*, *mstk0*, *mtas1*, *mtmr0* and *mwbt0* are assigned to the evaluation section. Moreover, the following subjects in the development section are to be used only as impostors: *fadg0*, *faks0*, *mjsw0* and *mmdb1*. Furthermore, the following subjects in the evaluation section are to be used only as impostors: *fedw0*, *felc0*, *mtmr0* and *mwbt0*.

The development section is to be used as follows. Non-impostor utterances from Session 1 are used to train a global model [3, 11]; the client models are then created by adapting the global model. True claimant and impostor accesses are simulated using utterances from Sessions 2 & 3, and are used to find the decision threshold (or parameters for an equivalent decision mechanism) which optimizes a given criterion (for example, the Equal Error Rate [4] or Decision Cost Function [1, 3]). In total there are 68 (17×4) true claimant tests and 272 ($17 \times 4 \times 4$) impostor tests.

The evaluation section is to be used as follows. The client models are created by adapting the global model trained using utterances from the development section (see above). True claimant and impostor accesses are simulated using utterances from Sessions 2 & 3; the performance is then calculated (for example, in terms of Decision Cost Function [1, 3]) using the decision threshold (or the decision mechanism) setup using the development section (see above). In total there are 72 (18×4) true claimant tests and 288 ($18 \times 4 \times 4$) impostor tests.

The meaning of the development and evaluation sections is then swapped and the above procedure for finding the performance figure is repeated. The resulting two performance figures are then averaged.

6 License

7 Acknowledgments

My thanks go to all the volunteers for their time, as well as Professor Kuldip K. Paliwal (Griffith University, Australia) and Dr Samy Bengio (IDIAP, Switzerland) for their valuable suggestions.

References

[1] S. Bengio and J. Mariéthoz, “Learning the Decision Function for Speaker Verification”, *Proc. International Conf. Acoustics, Speech and Signal Processing*, Salt Lake City, 2001.

[2] S. Bengio, J. Mariéthoz and S. Marcel, “Evaluation of Biometric Technology on XM2VTS”, *IDIAP Research Report* 01-21, Martigny, Switzerland, 2001.

[3] S. Bengio and J. Mariéthoz, “Comparison of Client Model Adaptation Schemes”, *IDIAP Research Report* 01-25, Martigny, Switzerland, 2001.

[4] G. R. Doddington, M. A. Przybycki, A. F. Martin and D. A. Reynolds, “The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective”, *Speech Communication*, Vol. 31, No. 2-3, 2000, pp. 225-254.

[5] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, John Wiley & Sons, USA, 2001.

[6] S. Furui, “Cepstral Analysis Technique for Automatic Speaker Verification”, *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 29, No. 2, 1981, pp. 254-272.

[7] S. Furui, “Recent Advances in Speaker Recognition”, *Pattern Recognition Letters*, Vol. 18, No. 9, 1997, pp. 859-872.

[8] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, “NTIMIT: A Phonetically Balanced, Continuous Speech Telephone Bandwidth Speech Database”, *Proc. International Conf. Acoustics, Speech and Signal Processing*, Albuquerque, 1990, Vol. 1, pp. 109-112.

[9] K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre, "XM2VTSDB: The Extended M2VTS Database", *Proc. Second International Conf. on Audio- and Video-based Biometric Person Authentication*, Washington D.C., 1999, pp. 72-77.

[10] S. Pigeon and L. Vandendorpe, "The M2VTS Multimodal Face Database (Release 1.00)", *Proc. First International Conf. on Audio- and Video-based Biometric Person Authentication*, Crans-Montana, Switzerland, 1997, pp. 403-409.

[11] D. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, Vol. 10, No. 1-3, 2000, pp. 19-41.

[12] C. Sanderson and K. K. Paliwal, "Polynomial Features for Robust Face Authentication", *Proc. International Conf. Image Processing*, Rochester, New York, 2002.

[13] C. Sanderson and K. K. Paliwal, "Likelihood Normalization for Face Authentication in Variable Recording Conditions", *Proc. International Conf. Image Processing*, Rochester, New York, 2002.

[14] C. Sanderson, "Automatic Person Verification Using Speech and Face Information", PhD Thesis, School of Microelectronic Engineering, Griffith University, Brisbane, Australia, 2002.