Challenge-based speaker recognition for mobile authentication

Mossab Baloul, Estelle Cherrier, Christophe Rosenberger

Challenge-based Speaker Recognition For Mobile Authentication

M. Baloul, E. Cherrier and C. Rosenberger

Université de Caen Basse-Normandie, UMR 6072 GREYC, F-14032 Caen, France
ENSICAEN, UMR 6072 GREYC, F-14050 Caen, France
CNRS, UMR 6072 GREYC, F-14032 Caen, France
Email: mossabini@hotmail.com, estelle.cherrier@ensicaen.fr, christophe.rosenberger@ensicaen.fr

Abstract

User authentication is a major trend to guarantee the security of electronic transactions when using mobile devices such as tablets or mobile phones. Biometrics is for us the only real user authentication method. In this article, we propose to realize a speaker recognition approach to achieve this goal. We use a challenge-based method to avoid the replay attack (especially if the impostor has recorded the user's voice). In this case, free text recognition is realized. Experimental results on the CMU database show very good results, while providing low computation times.

I. Introduction

The wide and recent development of smartphones and the correlated growing request to access online services (home banking, e-government, e-commerce…) has involved a need for mobile secure authentication. Among the existing solutions (static passwords, one time passwords, X509 certificates, coding tables…), challenge-based biometric authentication represents a promising proposal. Like any biometric system, a challenge-based biometric solution must meet essential requirements to address security and respect for privacy such as: confidentiality, unlinkability, resistance to replay attacks, revocability.

Within the biometric research field, challenge-based approaches are related to dynamic authentication that can be solved using a behavioral modality, such as mouse dynamics, keystroke dynamics, speaker recognition, etc… Behavioral biometrics has the advantage of being non-intrusive, in the sense that speaking, typing on a keyboard…is natural and simple for the user, therefore such modalities are globally well accepted.

Similarly to all biometric systems, challenge-based ones consist of two steps. The first step concerns the user enrolment: enrolment means first the capture of the biometric raw data, the features extraction to define a model (which is stored as a reference) of each genuine user and its storage (if the template meets some quality requirements). The second step called verification, used either for authentication or identification purposes, considering a challenge, must predict if the user has the expected behavior face to the challenge: as for example, type an unknown sentence on a keyboard or tell an unknown sentence… Since behavioral biometrics is involved, it must be difficult for an intruder to imitate the correct behavior.

Concerning mobile phones, some biometric sensors are already present in the object itself, providing them with inherent biometric abilities: we can mention the microphone, the webcam, the touch pad (and for some of them a fingerprint reader). Therefore, a challenge based on the way the mobile's owner speaks seems rather obvious and natural.

Challenge-based speaker recognition on a mobile phone belongs to the wide research field of text-independent speaker recognition. Indeed, to be authenticated, the mobile's owner will have to utter an unknown sentence or an unknown word, which is precisely text-independent speaker verification. Among the intense literature on this topic, we just refer the reader to the thorough survey paper [1] and the associated references. Using classical speaker recognition techniques to design an authentication system based on a biometric challenge on a mobile phone is not straightforward. Indeed, some constraints, inherent to the use of a mobile device, must be taken into account from the design step: the quality of the sound acquisition depends on the characteristics of the embedded microphone and the environment, the complexity of the embedded algorithms must be adapted to the capacity of the smartphone in terms of memory and processing power. The aim of this paper is twofold. First, how to find a simple solution that could be further embedded in a mobile, among the existing speaker recognition techniques? Second, what are the performances of the selected method applied to a suited database, in terms of EER, recognition rate and verification time?

The outline of the paper is the following: in Section II, we detail the different steps of a challenge based biometric speaker recognition for mobile devices. Both stages of enrollment and verification will be considered within the constraints inherent to the mobile context. Some methods of the literature will be presented and the most adapted one will be described. Section III presents the proposed method based on MFCC (Mel-frequency cepstral coefficients) characterization of the voice signal. Section IV is dedicated to the experimental protocol description and the obtained experimental results. At the end of the paper, the conclusion of our study and some perspectives will be given. We conclude and give some perspectives of this work.

II. Text-independent speaker recognition

The human voice is a complex information-bearing signal, depending on physical and behavioral characteristics. The raw speech signal, uttered by any person, is extremely rich in the sense that it involves high dimensional features. To perform efficient speaker recognition, one must reduce this complexity, while keeping sufficient information in the extracted feature vector. Some methods for speaker recognition have become popular, since few decades, which are gathered in the survey paper [1]. Here, we briefly recall the text-independent speaker recognition process, where five steps are considered.

- Signal acquisition
  Microphones and analog-digital converter are used to record and digitize the user's voice. At the end of this step, a numerical vector representing the uttered speech is available. The duration of speech recording depends on the desired accuracy.
- Speech signal preprocessing
  The speech signal is not a stationary signal since the vocal tract is continuously deformed and the model parameters are time-varying. But, it is generally admitted that these parameters are constant over sufficiently small time intervals. Classically, the signal is divided into frames of 25 milliseconds. This division into frames leads to discontinuities in the temporal domain, and inevitably to oscillations in the frequential domain. Among the possible solutions to avoid this phenomenon (see [2] for example), Hamming windows are applied. Besides, within the uttered text, silence zones can lead to performance degradation, so they must be removed. The reference [3] presents a voice activity detection (VAD) method based on realtime periodicity analysis, which enables silence removal. This method is also applied in [4]. In case of noisy signal, it can be filtered to reduce the noise level.
- Feature extraction
  Based on the speech signal registration and preprocessing, features are extracted to define a model corresponding to the user. Ideally, these features must be robust to intrinsic variability of the user's voice (due to stress, to disease), to noise and distorsion, to impersonation. The most widely employed methods involve short-term spectral features. We just cite two of them: MFCC (Mel-frequency cepstral coefficients) introduced by [5], and LPCC (linear predictive cepstral coefficients) proposed by [6], a detailed overview can be found in [1]. According to numerous studies, MFCC reveals to be more robust and efficient in practice.
- Speaker modeling
  Once these features have been extracted on each frame, the corresponding model or template design requires a training phase. We mention here the most popular techniques. GMM (Gaussian mixture model) [7] is a method based on a modeling of the statistical distribution of the extracted features. This method exhibits excellent performances, but is not suited to a challenge-based biometric system, owing to its computational cost. The VQ (vector quantization) method [8] is based on LBG algorithm [9], [10]. This process permits, after clustering, to describe a voice sample by a model vector having a predefined fixed size, whatever the initial length of the signal. Besides, the most recent method SVM (Support vector machine) [11] consists of binary classifiers, developed to allow the separation of complex data in large spaces. One SVM must be trained for

each genuine user.

- Speaker recognition
  These four previous steps correspond to the user enrolment phase. In the last recognition step, two problems can be considered: user authentication (the system must verify a claimed identity, through one vs. one comparison) or user identification (the system must check if the user is a genuine user, through one vs. multiple comparisons depending of the number of genuine users in the database). For GMM based modeling, the recognition relies on a likelihood estimation and the output is a probability. For VQ based modeling, the recognition test is classically performed through Euclidean distance computation. Whereas for SVM based modeling, the test phase uses the same process as the training phase. The reference [12] shows that the performances are at least as good as that of GMM based recognition. Notice that the acquisition conditions may be worse in this step than in the enrolment step, where the stored model must be of high quality.

  For GMM or VQ based modeling, the recognition test is classically performed through the Euclidean distance computation, with less parameters for the VQ. Whereas for SVM based modeling, the test phase uses the same process as the training phase. The reference [12] shows that the performances are at least as good as that of GMM based recognition. Notice that the acquisition conditions may be worse in this step than in the enrolment step, where the stored model must be of high quality.

  Many papers propose to use MFCC combined with SVM to perform speaker recognition, we mention just a few: in [13] for text-dependent speaker identification with neural networks, in [14] for text-dependent speaker verification, in the project [4] for a thorough implementation.

The main contribution of this paper is to analyze the performances of the previous algorithms, chosen among the most efficient of the literature, depending on different sets of parameters. We do not propose any new method, we combine existing methods to design a text-independent authentication process, applied to a realistic (concerning the acquisition conditions) database. The proposed analysis concerns a trade-off between the performance, and the computational cost. It is the first step in the design of an implementation on a mobile device. Now, we detail how the aforementioned speaker recognition techniques can be adapted to a mobile context.

III. Challenge-based speaker recognition

In a challenge-based biometric system, the enrolment phase is not different from that of any biometric system. Concerning speaker recognition, the user is asked to speak during a predefined time. Then, the preprocessing, the feature extraction and user modeling are performed to generate a template stored in a database. In the test mode and after extracting features, a distance between these parameters and the claimed model is calculated and compared to a given threshold, if it is a verification; and between these parameters and all models that exist in the database, if it is an identification.

In this paper, as in most of papers dealing with text-independent speaker recognition, we consider a MFCC based method owing to its robustness and better performances as a feature extraction method [1], [4] and VQ for modeling [15]. In figure III, we present the general diagram of a speaker recognition system in training and test mode.

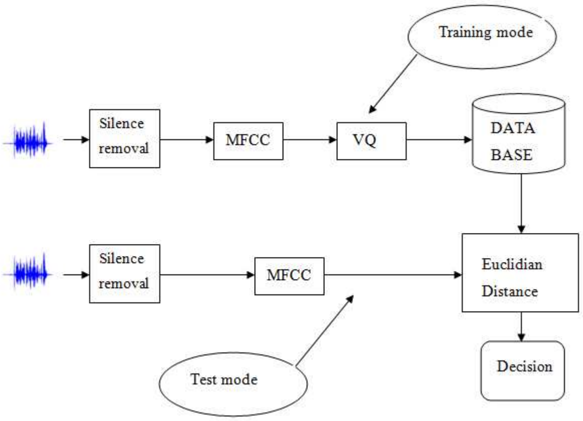

Fig. 1. General architecture of speaker recognition system

As mentioned in section II, the first step in speaker recognition is silence removal. In this paper, we resort to the simple method proposed in [16] to remove silence. The computation time of this approach is low, which is a very important property for mobile authentication. This method is based on the extraction of two particular audio features, namely signal energy and spectral centroid, defined below.

Let $x(n)$, $n \geq 0$ stand for one sentence of the database, and n the current discrete time. This signal $x(n)$ is divided into N frames of 50 milliseconds, denoted $x_i(n)$, for $i = 1, N$. For each frame $x_i(n)$, one defines:

- Signal Energy: the energy of the ith frame is computed as follows

$$E(i) = \frac{1}{N} \sum_{n=1}^{N} |x_i(n)|^2 \tag{1}$$

- Spectral Centroid: we can compute the spectral centroid C_i of the ith frame by the following formula

$$C_i = \frac{\sum_{k=1}^{N} (k+1) X_i(k)}{\sum_{n=1}^{N} X_i(k)} \tag{2}$$

where $X_i(k)$, $k = 1..., N$ stands for the ith discrete Fourier transform coefficient of the ith frame.

After computing these two feature sequences (Energy and Centroid), they will be compared to two thresholds T_E and T_C based on the energy sequences and spectral centroid sequences respectively. We describe the process to determine T_E, the same method is applied to determine T_C.

The histogram of the energy sequence is computed and then a smoothing filter is applied (a median filter). The threshold T_E is estimated as follows from the local maxima of the histogram as follows:

Let M_1 and M_2 denote the positions of the first and second local maxima respectively. Then, compute:

$$T_E = \frac{wM_1 + M_2}{w + 1} \tag{3}$$

where w is a user defined parameter.

The voiced frames are determined as the frames whose both feature values (i.e. Energy and Centroid) are larger then the two thresholds T_E and T_C respectively. After removing silence from the voice signal, the features extraction and modeling steps can be applied. In this paper, we use the MFCC and VQ algorithms detailed in the reference [17].

The figure 2 illustrates the computation of the MFCC coefficients, which is briefly detailed below.

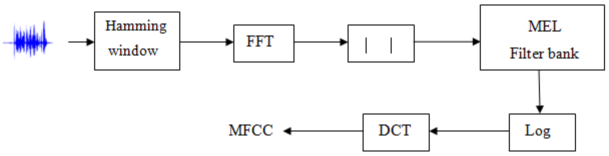

Fig. 2. Calculation process of MFCC coefficients

Consider again a particular sentence $x(n)$ as before. The voice signal is divided into small frames $x_i(n)$ of 256 samples with an overlap between them of 60 %. A Hamming window is applied to each frame:

$$y_i(n) = x_i(n) * w(n) \tag{4}$$

where $y_i(n)$ is the transformed signal, $x_i(n)$ is the considered frame and $W(n)$ is the Hamming window defined by:

$$W(n) = 0.54 - 0.46 cos(\frac{2\pi n}{256 - 1}) \tag{5}$$

for $0 \leq n \leq N - 1$.

The Fourier transform of each frame is computed, the next step is performed in the frequency domain. The human voice spectrum is not linearly distributed, therefore, we use a Mel scale filter bank to represent the wide spectrum. A given frequency f in Hz can be converted into the Mel scale [18]:

$$MEL(f) = 2595 * log_{10}(1 + \frac{f}{700}) \tag{6}$$

In general, 20 Mel filters are required for high accuracy. We apply after a logarithmic compression and a discrete cosine transform. Finally, the discrete amplitudes of the resulting cepstrum are called the MFCCs coefficients [5].

The resulting MFFC coefficients of each sentence $x(n)$ are 20 dimensional vectors, each vector will be represented by a given number of centroids (between 8 and 256), resulting in a vector template of fixed size modeling each user. This step is called the Vector Quantization (VQ), it is done by the LBG algorithm [15].

One advantage of using VQ is to reduce the computational cost. The obtained centroids are used to model the user. At the enrolment step, for each user, a specific model is calculated and stored in the database.

At the verification step, after the extraction of the query MFCC coefficients, we compute the Euclidean distance between these parameters and the model of the claimed reference; the obtained distance is compared to a given threshold.

Let $MFCC(n,p)$ be the MFCC coefficients of a given user and $VQ(n,q)$ the query reference model, $p >> q$:

$$MFCC = \begin{pmatrix} MFCC_{11} & MFCC_{12} & ... & MFCC_{1p} \\ MFCC_{21} & MFCC_{22} & ... & MFCC_{2p} \\ . & & & \\ . & & & \\ . & & & \\ MFCC_{n1} & MFCC_{n2} & ... & MFCC_{np} \end{pmatrix} \tag{7}$$

$$VQ = \begin{pmatrix} VQ_{11} & VQ_{12} & ... & VQ_{1q} \\ VQ_{21} & VQ_{22} & ... & VQ_{2q} \\ . & & & \\ . & & & \\ . & & & \\ VQ_{n1} & VQ_{n2} & ... & VQ_{nq} \end{pmatrix} \tag{8}$$

To compute the Euclidean distance between the two matrices $MFCC$ and VQ, we proceed as follows.

For each column $MFCC_j$, $j = 1, p$, we calculate the Euclidean distance between this column and the nearest column VQ_k, $k = 1, q$.

The considered distance is the sum of the p resulting distances, following:

$$ED_{(MFCC,VQ)} = \sum_{j=1}^{p} \min_{1 \leq k \leq q} \{dist(MFCC_j, VQ_k)\} \tag{9}$$

where :

$$dist(MFCC_j, VQ_k) = \sqrt{\sum_{l=1}^{n} (MFCC_{l,j} - VQ_{l,k})^2} \tag{10}$$

where $MFCC_{l,j}$ (respectively $VQ_{l,k}$) is the coefficient of the $MFCC$ matrix (resp. the VQ matrix) at row l and column j (resp. row l and column k).

The final decision is the result of all this process, it depends on an operational threshold: if the distance $ED_{(MFCC,VQ)}$ is lower than this threshold, the user is authenticated by the system, otherwise it is rejected.

In the next section, we characterize the proposed method in terms of performance and computation time. We also analyze the impact of parameters on efficiency such as the number of centroids to consider for the quantization step or the number of samples to generate the model of the user.

IV. Experimental results

In this paper, we consider the PDA database (PDAm data set) [19] proposed by CMU (Carnegie Mellon University). It consists of voice signals collected by a PDA device. 50 sentences of about 4 to 8 seconds are uttered by 16 users. The users work at CMU, they are native speakers of American English. The voice samples are recorded at 44.1kHz sampling rate. The original data was then downsampled to both 16kHz and 11.025kHz, see [19] for more details.

In this paper, we quantify the performance of the proposed method as follows:

- The number of sentences used in the training step are varied (1, 10, 20, 30 and 40 sentences among 50). For each value, the number of centroids used in the VQ method are also varied.
- We compute the ROC (Receiver Operating Curve) curve that gives the performance behavior of the biometric system for any value of the decision threshold (for verification purpose).
- The performance is also evaluated through the computation of the EER (Equal Error Rate).
- The second performance criterion is the recognition accuracy.
- We intend to determine the best tradeoff between a low EER, a small number of centroids which influence the memory space and a small number of sentences for the enrolment step which has an impact on the execution time.
- We propose to evaluate the time necessary for the enrolment step, depending on the previous chosen parameters, on a PC, since this step could be performed on a server side.

In table I, we present the EER value of the biometric system by varying the number of sentences and the number of centroids used in the VQ modeling. We can see that the more centroids we use, the better is the performance. As the number of centroids used for the enrolment step has an impact of the computation time, we try to find a tradeoff between efficiency and computation time.

We obtain the best EER = 0.83 with 30 sentences (which is equivalent to about 3 minutes of recorded voice) and 64 centroids for the VQ modeling. This performance is interesting for a low cost biometric solution. In order to avoid the replay attack, the voice can be analysed in order to match the challenge.

Figure IV shows the corresponding ROC curve. We can see that for the FRR value $FAR = 10^{-4}$ equals 20% which is not bad for a low cost solution. In table II, we present the recognition rate of the system by varying the number of sentences and the number of centroids used for the VQ modeling. These results are

	Number of sentences				
	1	10	20	30	40
VQ 8	12.52	8.14	6.68	6.68	7.72
VQ 16	6.68	6.47	6.05	5.21	1.46
VQ 32	6.05	6.05	5.21	1.25	1.25
VQ 64	6.26	6.47	1.46	0.83	1.04
VQ 128	6.05	6.26	1.46	0.83	0.83
VQ 256	6.05	6.05	1.46	0.83	0.83

TABLE I. EER for different numbers of sentences and centroids

satisfying as the False Acceptance Rate is in general low and it is very easy and fast to ask the user to make another capture for the verification step.

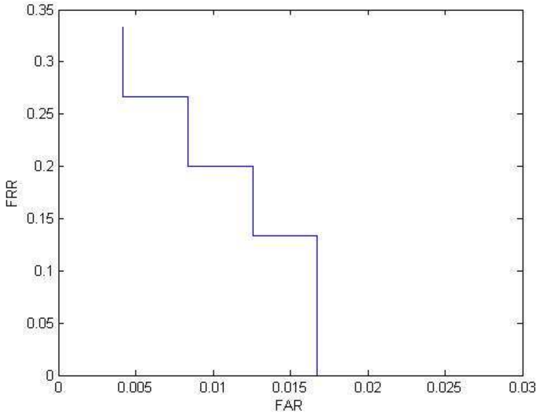

Fig. 3. ROC curve for 30 sentences and 64 centroids

	Number of sentences				
	1	10	20	30	40
VQ 8	91.58	94.06	93.75	94.69	96.88
VQ 16	92.47	92.81	93.54	94.38	95.62
VQ 32	92.86	93.28	93.54	94.38	95.62
VQ 64	92.98	93.75	93.54	94.38	95.62
VQ 128	92.6	93.59	93.54	94.38	95.62
VQ 256	92.86	93.75	93.33	94.38	95.62

TABLE II. Authentication rate for different numbers of sentences and centroids

Now, bearing in mind the initial purpose of mobile implementation, we intend to estimate the computation time of both steps: enrolment and verification. The enrolment step will probably be done on a PC used as server, so this estimation has been performed with ©Matlab, with the selected values for the parameters (number of sentences and number of centroids). We proceed as follows: to estimate the enrolment time, consider 30 seconds of the recorded voice and design a model of this voice signal with 64 centroids. It takes 35.6 seconds, so for 3 minutes of speech signal it will take about 3 minutes and 34 seconds in ©Matlab environment. With a C implementation, we can expect to decrease by ten the computation time.

For the verification time, a sample of 5 seconds of voice signal is selected for the verification process. It takes 2.53 seconds of time processing in ©Matlab. Even if it is difficult to compare the computation time between ©Matlab on laptop and a mobile phone, we think that this computation time is a good estimate on what we could achieve on a mobile device.

V. Conclusion and perspectives

We proposed in this article a free-text speaker recognition method. The features we used are Mel-frequency cepstral coefficients. The vector quantization allows to handle fixed-size feature vectors. We optimized the processing chain in order to have a good tradeoff between efficiency and computation time. Recognition results on the CMU database (that represents operational conditions) are satisfying with a EER value equal to 0.83.

Perspectives of this work are to implement on a mobile phone the proposed method to realize an off-line user authentication.

References

[1] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Communication*, vol. 52, pp. 12–40, 2012.

[2] F. Harris, “On the use of windows for harmonic analysis with the discrete fourier transform,” *Proceedings of the IEEE*, vol. 1, pp. 51–83, 1978.

[3] V. Hautam, M. Tuononen, T. Niemi-Laitinen, and P. Franti, “Improving speaker verification by periodicity based voice activity detection,” in *Proceedings of the 12th International Conference on Speech and Computer*, 2007.

[4] P. Fränti, J. Saastamoinen, I. Kärkkäinen, T. Kinnunen, V. Hautamäki, and I. Sidoroff, “Developing speaker recognition system: from prototype to practical application,” in *Forensics in Telecommunications, Information and Multimedia*. Springer Berlin Heidelberg, 2009.

[5] S. Davis and P. Mermelstein, “Comparison of parametric representations for mmonosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 28, pp. 357–366, 1980.

[6] X. Huang, A. Acero, and H. Hon, *Spoken language processing: a guide to theory, algorithm, and system development*. Prentice-Hall, New Jersey, 2001.

[7] D. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Transactions on Speech Audio Processing*, vol. 3, pp. 72–83, 1995.

[8] F. Soong, A. Rosenberg, L. Rabiner, and B.-H. Juang, "A vector quantization approach to speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1985.

[9] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. COM-28, pp. 84–95, 1980.

[10] D. Burton, "Text-dependent speaker verification using vector quantization source coding," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 35, pp. 133–143, 1987.

[11] V. Vapnik, *Statistical learning theory*. Wiley, 1998.

[12] D. Matrouf, J.-F. Bonastre, C. Fredouille, A. Larcher, S. Mezaache, M. McLaren, and F. Huenupan, "Lia gmm-svm system description : Nist sre08," in *NIST Speaker Recognition Evaluation Workshop*, 2008.

[13] S. M. Kamruzzaman, A. N. M. R. Karim, M. S. Islam, and M. E. Haque, "Speaker Identification using MFCC-Domain Support Vector Machine," *International Journal of Electrical and Power Engineering*, vol. 1, pp. 274–278, 2007.

[14] S.-H. Chen and Y.-R. Luo, "Speaker verification using mfcc and support vector machine," in *IMECS International Multi-Conference of Engineers and Computer Scientists*, 2009.

[15] A. Kabir and S. Ahsan, "Vector quantization in text dependent automatic speaker recognition using mel-frequency cepstrum coefficient," in *Proceedings of 6 th WSEAS International Conference on Circuits, Systems, Electronics, Control and Signal Processing, Cairo, Egypt*, 2007.

[16] T. Giannakopoulos, "A method for silence removal and segmentation of speech signals, implemented in matlab," *Department of Informatics and Telecommunications, University of Athens, Greece , Computational Intelligence Laboratory (CIL) , Insititute of Informatics and Telecommunications (IIT) , NCSR DEMOKRITOS, Greece*.

[17] V. Velisavljevic, C. Cornaz, and U. Hunkeler, "Mini-project: : An automatic speaker recognition system," EPFL, Tech. Rep., 2003.

[18] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient and dynamic time warping techniques," *Computing Research Repository*, 2010.

[19] Y. Obuchi, "PDA speech database, carnegie mellon university, available at http://www.speech.cs.cmu.edu/databases/pda/index.html."