IMPROVING SPEAKER VERIFICATION PERFORMANCE UNDER SPOOFING ATTACKSBY FUSION OF DIFFERENT OPERATIONAL MODES

Saeid Safavi, Hock Gan, and Iosif Mporas

Information Engineering and Processing Architectures Group, ECE Division
School of Engineering and Technology, University of Hertfordshire
College Lane Campus, Hatfield, AL10 9AB, UK
{s.safavi, h.c.gan, i.mporas}@herts.ac.uk

ABSTRACT

In this paper, we propose a methodology for the fusion of different modes of speaker verification (SV) operation (fixed-passphrase, text-dependent and text-independent mode), using regression fusion models. The experimental results with and without spoofing attack conditions and using different single mode speaker verification engines, GMM-UBM, HMM-UBM and i-vector, indicated improvement in all the experiments. The 6.75 % in terms of EER is achieved as the best speaker verification performance, when using fusion of scores from three modes of operation of HMM-UBM based speaker verification systems. Relative improvement of 22.32% achieved compare to the best performing single mode engine.

***Index Terms*—**Automatic speaker verification, spoofing attack, anti-spoofing, regression fusion, i-vector, HMM-UBM, GMM-UBM.

1. INTRODUCTION

Biometric authentication has significantly advanced over the last decade, with exceptional results in the speaker verification task. Specifically, speaker verification was well established by the probabilistic methodology proposed by Reynolds [1]and has further been improved by more sophisticated approaches[2, 3, 4, 5]which are mainly based on algorithms from the areas of machine learning and linear algebra. These methodologies are regularly evaluated by well-known scientific challenges and evaluations such as the NIST SRE (Speaker Recognition Evaluation)[6, 7]the AVSpoof database evaluations[8]and the RedDots challenge[9].

Although a number of different approaches and architectures have been proposed for speaker verification, three major methods can be detected: the GMM-UBM[10], the HMM-UBM[11] and the recently proposed i-vectors[2] approach. In the GMM-UBM method a Gaussian mixture model (GMM) is used to train a Universal Background Model (UBM) using recordings from a large number of speakers and is then adapted to the target speaker's enrolment recordings, usually using mean-only adaptation.In the HMM-UBM method, a similar methodology to the GMM-UBM is adopted with the difference that here the UBM and the target speakerare built by hidden Markov models, thus able to model temporal information. In the case of i-vectors, the mean of the Gaussian distributions of the GMM-UBM models are constructing a super-vector, which is a descriptor of the whole voice input and using joint factor analysis is split into channel and speaker vectors. The i-vectors approach achieves state-of-the-art performance in cases where significant amount of training datais available[12].

Despite having more than 30 years of research effort in the area of speaker verification, voice biometric protection against spoofing is still open for improvement since most speaker verification approaches are vulnerable to spoofing attacks. There are four major types of spoofing attacks, namely the impersonation, the audio replay, the speech synthesis and the voice conversion attack[13].Comparison of spectrograms between genuine and impersonated voice samples have shown that the formants do not quite match each other.The drawback of speech synthesis and voice conversion based spoofing attacks is the phase[14] and prosody (F_0 statistics)[15] information, which are often not speech-like. In audio replay attacks, countermeasures based on uncharacteristic similarity between recorded inputs[16] and dissimilarity due to the specific environmental characteristics of replay attacks[17] have been proposed. However, when the recording and replay acoustic environment conditions are similarthe detection of audio replay spoofing attack is difficult and thus the voice biometric applications become vulnerable to attacks[13].

In this paper, we propose a speaker verification architecture, which relies on the combination of text-dependent and text-independent voice inputs and has applications in real-life sensitive voice biometrics applications, for example Banking.

2. SPEAKER VERIFICATION USING FUSION OF OPERATIONAL MODES

In real-life voice authentication interfaces,three modes of operation can be found, namely the fixed-passphrase, thetext-dependent and the text-independent mode. In the fixed-passphrase mode, the user is spelling a password or passphrase, which has been predefined during the enrolment phase. The target speaker model has been trained using enrolment voice samples of the fixed passphrase and thus, in general, this mode of operation achieves low equal error rates (EERs). In the case of text-dependent mode, the user is prompted with a text message selected from a closed set of a few (usually 10-20) utterances. In this mode, the acoustic model of the speaker is trained from enrolment voice samples of these utterances and the EER is slightly higher compared with the EER of the fixed passphrase mode of operation. In the text-independent case, the user is prompted with randomly selected utterances (e.g. random word sequences or sentences produced by natural language generation engines).In this mode, the speaker models are trained with voice samples of the speaker that are text-independent and the performance (in terms of EER) is significantly worse compared with the EERs of the fixed passphrase mode.

When comparing EERs,the text-independent mode is more robust against spoofing attacks, since it is quite improbable that the same text prompted by the systemhas been recorded for the audio replay attack, while for the case of speech synthesis and voice conversion the produced synthetic speech is of a lower quality. The latter case is ineffective despite being able to respond to the impromptu demands of the system.Conversely, the fixed-passphrase mode is highly vulnerable to audio replay attacks because it is relatively easy to produce high quality synthetic speech for known phrases.

Obviously, there is a trade-off between EER performance and anti-spoofing robustness, so an appropriate combination of these three modes of operation would result in a voice biometric interface that is robust against spoofing attacks while retaining low EERs. The proposed architecture for fusion of different modes of speaker verification operation is illustrated in Fig. 1.

As can be seen in Fig. 1, the user is asked to provide a fixed passphrase voice input, afterwards a text-dependent input and finally a text-independent input. Each of the three voice samples is processed by a mode-specific acoustic model as described above. The three produced speaker verification scores are concatenated to a 3 dimensional feature vector and fused by a regression model, in order to produce the final speaker verification score.

Let us denote the three voice inputs as v_a, v_b and v_c, as shown in Fig. 1. Each voice input will be processed by a mode-specific speaker model,M_a,M_b, M_c and produce one speaker verification score for each voice input, i.e. s_a, s_b and s_c ,respectively. The three scores are concatenated to a global vector $s = [s_a, s_b, s_c]^T$ with $s \in \Re^3$. A regression model, f, is used to fuse the different modes' scores to produce a final speaker verification score, $d = f(s)$, with $d \in \Re$.

The combination of the scores of different modes of operation by the regression model will exploit the complementary

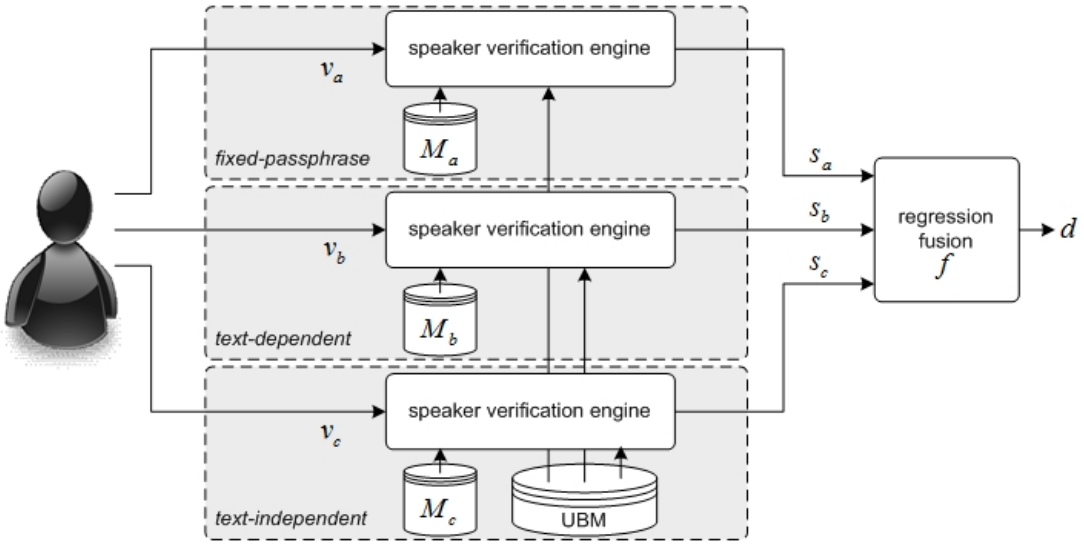

Fig.1: Block diagram of the proposed methodology for speaker verification using fusion of different modes of operation.

information between them and result in an improved voice authentication.

3. EXPERIMENTAL SETUP

The proposed architecture for speaker verification was evaluated both in terms of improvement of speaker verification performance based on the EER criterion and in terms of robustness against spoofing attacks.

3.1. Data description

In the present evaluation, we used the AVSpoof [8] dataset. Thedataset consists of voice recordings with the sampling frequency at 16 kHz and a 16-bit resolution analysis. AVSpoof includes voice recordings from 31 male and 13 female speakers and was intentionally designed for evaluation of speaker verification systems against spoofing attacks such as audio replay attack, speech synthesis and voice conversion. In this work, we use the audio replay case as the spoofing attack, which is the hardest attack to tackle, although it is quite improbable that it appearsin the text-independent mode of operation. The setup of the three modes of operation (Section 2)from the AVSpoof data was performed within the Octave Project [18]. For all modes of speaker verification operation, the standard development and evaluation setup of the database was followed.
In order to use the data, we need to divide the data of each gender into three subsets, but as the number of female speakers is small, we only considered male speakers for the experiments in this report. All the data from the 31 male speakers were divided into two sections, Development and Evaluation sets, and each of these sets contain training and testing subsets. The Development and Evaluation set had data for 10 (m014 to m030) and 11 (m033 to m044) speakers, respectively. Table 1 contains statistics for three modes of speaker verification operation for both Development and Evaluation sets.Fixed-pass phrase protocols were designed such that each target speaker had five different phrase-specific models created separately for five pass phrases. For training only, one speech segment was used for each enrolment. During the test phase each test phrase wasscored against the trained segment model that containedthe same text from the same speaker (referred to as target trials) and different speaker (non-target trials).For mode (B) each target model was enrolled with data from five pass phrases of session 1, while in Mode (C) each target model was enrolled with first five sentences of free speech. But for both modes, B and C, are sharing a unique test trial list. The trial statistics for mode B and C is tabulated in Table 1.

Table 1: Trial statistics for three modes of operation of speaker verification, mode A, B and C.

Statistics	***Data sets***			
	Mode A		***Mode B & C***	
	Eval.	***Dev.***	***Eval.***	***Dev.***
Target models	55	50	11	10
Target trials (genuine)	494	450	494	450
Non-target trials (replay spoofed)	600	600	600	600
Non-target trials (zero-effort)	4940	4050	4940	4050

3.2. Single-mode speaker verification engines

To demonstrate the validity of the proposed fusion methodology we relied on three of the most widely used speaker verification modeling approaches for a baseline, namely the GMM-UBM [1, 19, 20], the HMM-UBM and the i-vectors approach. With respect to the GMM-UBM approach, we used a mixture of 256 Gaussian distributions and mean-only adaptation from the UBM model. For the HMM-UBM model, we used 14 state models with eight mixtures per state and left-to-right architecture [21]. With respect to the i-vectors approach, the T matrix was trained using the same data as UBM and it was trained for a 400-dimensional total variability space, full description of i-vector framework could be found in [22].
In all three approaches speaker models were trained using 57 dimensional Mel Frequency Cepstral Coefficients (MFCC) feature vectors consisting of static $C_1 - C_{19}$ cepstra, with Δ and $\Delta\Delta$features extracted from the speech signal using 10 ms shift and a 20 ms hamming window. An energy based voice activity detector is applied to remove the less energized frames. In order to decrease the channel effect and remove convolutive noise, cepstral mean and variance normalization and RASTA filtration [23] were applied on feature vectors.

3.3. Speaker verification fusion

For the fusion of the three modes of operation, we relied on regression models. Specifically, we used the linear regression model[24], the multilayer perceptron model[25] with one hidden layer and the support vector regression[26] algorithm with a RBF kernel withC=25007 and gamma=0.01. For all regression fusion algorithms, we used the WEKA [27] toolkit implementation. In order to make sure that the fusion approach does not biased to the speaker's information, for training the fusion parameters the scores from development speakers were used and for testing the fusion approach the scores from the Evaluation set were used.

4. EXPERIMENTAL RESULTS

The proposed methodology for fusion of speaker verification modes of operation was evaluated using the experimental setup described in the previous section. For both single mode and fusion approach an identical protocol from the evaluation set were used to be able to compare their performances. We evaluate the method in terms of speaker verification performance, i.e. EER, while exposed to audio replay attacks.

The detection error tradeoff (DET) curves for the HMM-UBM, the GMM-UBM and the i-vector system are shown in Fig. 2, Fig. 3 and Fig. 4, respectively. For direct comparison, the single-mode speaker verification performance is also illustrated, indicated as A for fixed passphrase, B for text-dependent and C for text-independent mode.

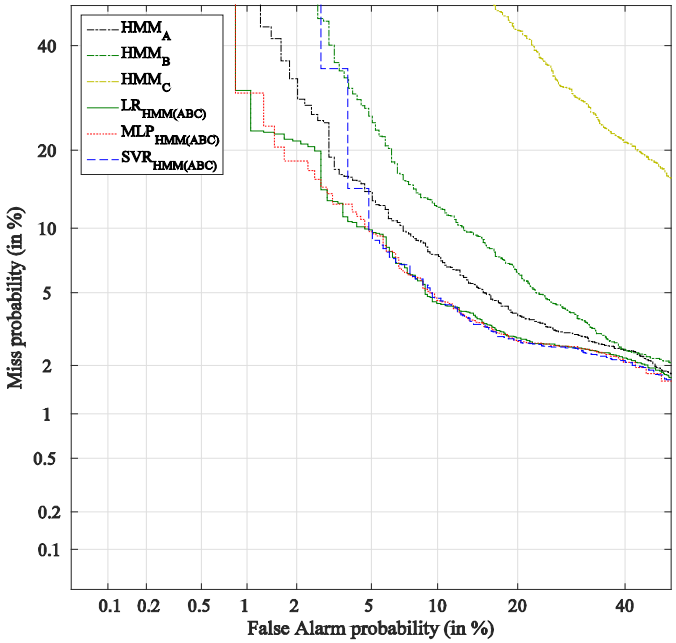

Fig.2: Speaker verification performance for fusion of modes of operation using HMM-UBM single-mode engines.

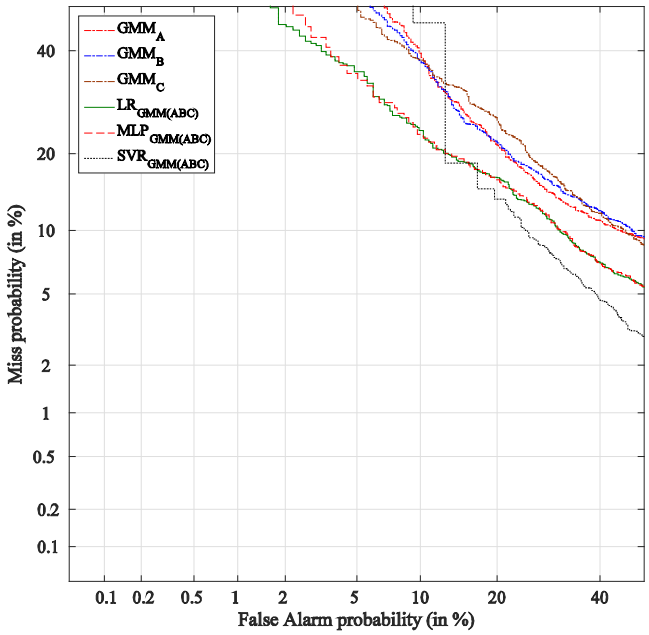

Fig.3: Speaker verification performance for fusion of modes of operation using GMM-UBM single-mode engines.

As can be seen in all the figures, the fusion of different modes of operation shows significant improvement in the speaker verification performance, regardless of the use of the GMM-UBM, the HMM-UBM or the i-vector approach. This is due to the complementary information between the different modes of operation.All three of the regression fusion algorithms outperformed the single mode-speakerverification systems.

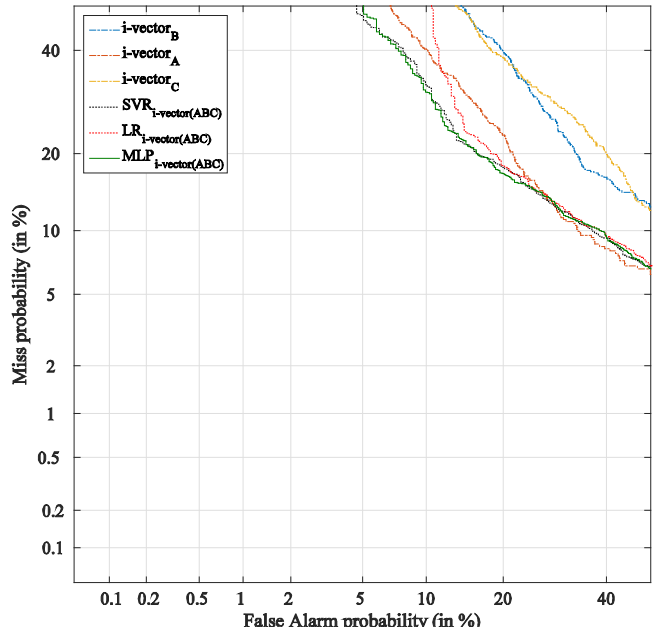

Fig.4: Speaker verification performance for fusion of modes of operation using i-vector single-mode engines.

The performance of each of the fusion algorithms is tabulated in Table 2.As can be seen in Table 2, the neural network regression model achieved the best performance in terms of EER, followed by the support-vector regression model.

Table 2: Speaker verificationequal error rate (EER %) of the evaluated regression fusion algorithms.

	A	*B*	*C*	*LR*	*MLP*	*SVR*
GMM-UBM	20.73	20.86	**22.78**	17.50	16.96	17.34
HMM-UBM	**8.69**	**11.31**	29.70	6.94	**6.75**	6.89
i-vector	20.91	27.32	29.08	19.23	18.42	18.62

5. CONCLUSION

We presented a methodology for the fusion of different modes of speaker verification operation, such as the fixed-passphrase, the text-dependent and the text-independent, using regression models. The experimental results indicated the validity of our method under the extreme case audio replay attacks independently of the approach used for the single mode speaker verification. We speculate that the proposed fusion approach is applicable to robust real-life voice authentication interfaces.

6. ACKNOWLEDGEMENT

This work was partially supported by the H2020 OCTAVE Project entitled “Objective Control for Talker Verification” funded by the EC with Grand Agreement number 647850. The authors would like to thank Dr Md Sahidullah, Dr Achintya Sarkar, and Dr Tomi Kinnunen for their support in this work.

7. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, no. 1, pp. 19-41, 2000.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta and P. Dumouchel, "A Study of Inter-Speaker Variability in Speaker Verification," *IEEE TRANS. AUDIO SPEECH AND LANGUAGE PROCESSING,* vol. 16, no. 5, pp. 980-988, 2008.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 4, pp. 788-798, 2011.

[4] T. Ganchev, M. Siafarikas, I. Mporas and T. Stoyanova, "Wavelet basis selection for enhanced speech parametrization in speaker verification," *International Journal of Speech Technology,* vol. 17, no. 1, pp. 27-36, 2014.

[5] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication,* vol. 73, pp. 1-13, 2015.

[6] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation.," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[7] R. Saeidi, K. A. Lee, T. Kinnunen, T. Hasan, B. Fauve, P. M. Bousquet, E. Khoury, P. L. Sordo Martinez, J. M. K. Kua, C. H. You, H. Sun and et al., "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," *Proceedings of Interspeech 2013,* pp. 1986-1990, 2013.

[8] S. K. Ergünay, E. Khoury, A. Lazaridis and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on,* pp. 1-6, 2015.

[9] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma and H. Li, "The RedDots Data Collection for Speaker Recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication,* vol. 17, no. 1, pp. 91-108, 1995.

[11] C. H. Lee and J. L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters," *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on,* vol. 2, 1993.

[12] S. S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke and L. Ferrer, "The SRI NIST 2008 speaker recognition evaluation system," *2009 IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 4205-4208, 2009.

[13] W. Zhizheng, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication,* vol. 66, pp. 130-153, 2015.

[14] P. L. De Leon, M. Pucher, J. Yamagishi and I. Hernaez, "Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 8, pp. 2280-2290, 2012.

[15] A. Ogihara, U. Hitoshi and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE transactions on fundamentals of electronics, communications and computer sciences,* vol. 88, no. 1, pp. 280-286, 2005.

[16] W. Shang and M. Stevenson, "Score normalization in playback attack detection," *2010 IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 1678-1681, 2010.

[17] Z. Wang, G. Wei and Q. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," *Machine*

Learning and Cybernetics (ICMLC), 2011 International Conference on, pp. 1708-1713, 2011.

[18] R. E. A. -. E. Commission, "OCTAVE Project – Objective Control of TAlker VErication:," 01 06 2015. [Online]. Available: https://www.octave-project.eu/. [Accessed 15 09 2016].

[19] S. Safavi, A. Hanani, M. Russell, P. Jancovic and M. Carey, "Contrasting the effects of different frequency bands on speaker and accent identification," *IEEE Signal Processing Letters,* vol. 19, no. 12, p. 829–832, 2012.

[20] S. Safavi, M. Najafian, A. Hanani, M. Russell and P. Jancovic, "Speaker recognition for children's speech," *Interspeech, pp. 1836-1839,* 2012.

[21] A. Sarkar, Z. H. Tan, "Text dependent speaker verification using unsupervised HMM-UBM and temporal GMM-UBM" *interspeech*, pp. 362-366, 2016.

[22] S. Safavi, M. Russell, P. Jancovic, "Identification of age-group from children's speech by computers and humans." *Interspeech,* pp. 243-247, 2014.

[23] H. Hermansky and N. Morgan, "Rasta processing of speech", *IEEE trans. on speech and Audio Processing*, vol. 2, pp. 578-589, 1994.

[24] C. M. Bishop, Pattern recognition and Machine Learning, Cambridge: Springer, 2006.

[25] W. S. Sarle, "Neural Networks and Statistical Models," in *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Dallas, 1994.

[26] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing,* vol. 14, no. 3, pp. 199-122, 2004.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, no. 1, pp. 10-18, 2009.