Владислав Александрович Сухмель — аспирант; Санкт-Петербургский государственный университет, кафедра компьютерного моделирования и многопроцессорных систем; E-mail: sukhmel@apmath.spbu.ru

Алексей Владимирович Шолохов — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: sholokhovalexey@gmail.com

Тимур Сахиевич Пеховский — канд. физ.-мат. наук; ООО „ЦРТ-инновации“, Санкт-Петербург; ведущий научный сотрудник; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: tim@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 621.391.037.372

В. Л. Щемелинин, К. К. Симончик

ИССЛЕДОВАНИЕ УСТОЙЧИВОСТИ ГОЛОСОВОЙ ВЕРИФИКАЦИИ К АТАКАМ, ИСПОЛЬЗУЮЩИМ СИСТЕМУ СИНТЕЗА

Проанализирована устойчивость современных методов верификации к взлому при помощи гибридной системы синтеза речи на основе технологий Unit Selection и скрытых марковских моделей. Представлен метод взлома, обеспечивающий достижение ошибки ложного пропуска в 98—100 % случаев при большом объеме обучающей базы; метод может быть автоматизирован при сопряжении с автоматической системой распознавания речи.

Ключевые слова: *спуфинг, синтез речи, распознавание диктора.*

Введение. Системы верификации дикторов по голосу широко используются в криминалистических экспертизах, системах контроля доступа, банковской сфере, а также Интернете. Основные задачи подобных систем — повышение удобства использования и защита от несанкционированного доступа [1]. Соревнования NIST SRE 2012 [2] показали, что преобладают системы, основанные на представлении модели голоса диктора в пространстве полной изменчивости (*total variability*). Однако, как показывают исследования, современные системы верификации неустойчивы к спуфингу [3] с помощью автоматического синтеза голоса.

В настоящей работе исследована зависимость надежности системы верификации от объема речевого материала для обучения системы синтеза.

Система голосовой верификации. Предлагаемый метод заключается в использовании смесей гауссовых распределений (Gaussian Mixture Models, GMM) для моделирования голоса диктора, а затем их редукции до так называемого *i*-вектора в низкоразмерном пространстве полной изменчивости.

В работе использованы система текстозависимой верификации дикторов на базе *i*-векторов [4, 5], а также специальный модуль препроцессинга, включающий энергетический детектор речи и детектор клиппированных сигналов [6] для их отбраковки. В качестве речевых признаков выступали векторы мел-частотных кепстральных коэффициентов (Mel-frequency Cepstrum Coefficients, MFCC), их производных первого и второго порядка (39 элементов). Длина каждого речевого кадра для вычисления MFCC составляла 22 мс со сдвигом 11 мс. Для компенсации эффекта Гиббса использовалось взвешивание сигнала окном Хем-

минга. Эффекты канальных искажений на уровне признаков компенсировались путем вычитания кепстрального среднего (*cepstral mean subtraction*). Выравнивание признаков (*feature warping*) [7] не применялось, поскольку длительность речевого сигнала была мала.

На этапе моделирования голоса диктора использовалась гендеронезависимая универсальная фоновая модель (Universal Background Model, UBM), представленная 512-компонентной смесью гауссовых распределений. Обучение UBM производилось с помощью стандартного EM-алгоритма на телефонной части речевых баз данных NIST SRE 1988—2010 [8, 9]. Для ускорения вычислений использовалась диагональная ковариационная матрица UBM. Общее число дикторов в обучающих базах данных — около 4000. Модуль оценки *i*-вектора (как и модуль линейного дискриминантного анализа) был обучен более чем на 60 000 телефонных и микрофонных записях из тех же речевых баз данных.

Модель GMM в низкоразмерном пространстве полной изменчивости представляется следующим выражением:

$$\mu = m + T\omega + \varepsilon,$$

где μ — супервектор параметров GMM модели диктора, m — супервектор параметров UBM, T — матрица, задающая базис в редуцированном пространстве признаков, ω — *i*-вектор в редуцированном пространстве признаков, $\omega \in N(0,1)$, ε — вектор ошибки.

Метод взлома системы верификации. Известны различные способы спуфинга (от англ. *spoofing* — получение доступа обманным путем). Например, в работе [10] описываются способы на основе воспроизведения записи голоса, манипуляции с записью голоса, прикрытия рта носовым платком или закрытия носа рукой. В настоящей работе для спуфинга используется гибридный метод синтеза на основе Unit Selection и скрытых марковских моделей (Hidden Markov Model, HMM) [11].

Метод взлома предполагает создание синтезированного голоса пользователя системы верификации. Для обучения системы синтеза используется предварительно записанная спонтанная речь пользователя. На этапе текстозависимой верификации при помощи синтезированного голоса и перехваченного парольного текста создается парольная фраза, используемая далее для попытки верификации. Детальная схема атаки представлена на рис. 1.

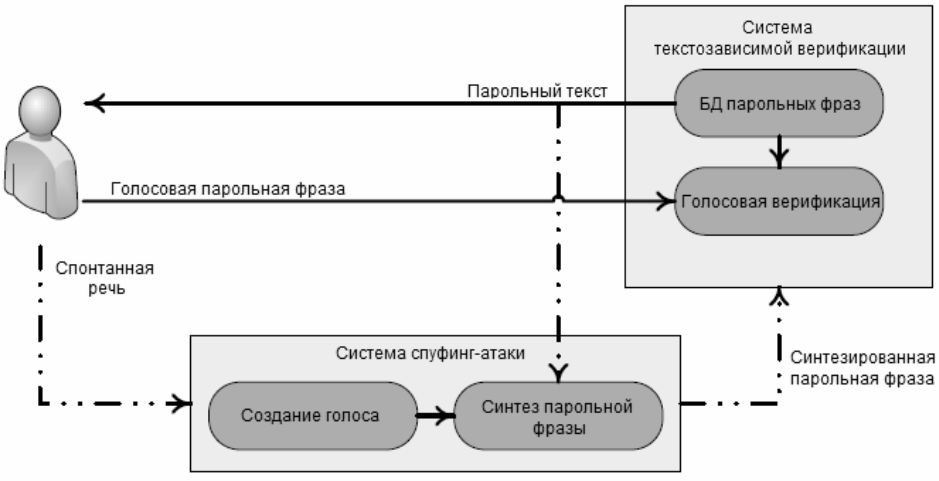

Рис. 1

Система синтеза голоса. Для моделирования спуфинг-атаки была использована система голосового синтеза, разработанная в ООО „ЦРТ“ [12]. В ней использованы два наиболее популярных подхода:

1) алгоритм Unit Selection (выбор речевых элементов), позволяющий достичь максимальной естественности синтезированной речи, при условии корректно отсегментированной на разных уровнях сбалансированной речевой базы данных большого объема;

2) статистические модели (HMM-синтез) позволяют легко модифицировать характеристики голоса с помощью адаптации/интерполяции дикторов. Речь, полученная на основе HMM-технологии, на слух менее естественна, однако в ней отсутствуют резкие, не обусловленные контекстом перепады по частоте и энергии, обычно присущие конкатенативному синтезу. Кроме того, применение HMM-синтеза позволяет разрабатывать новый голос за гораздо меньшее время, а также требует значительно меньше памяти для хранения речевой базы.

Речевой корпус. Для проведения экспериментов была использована речевая база русского языка, содержащая 7 различных дикторов (2 мужчин и 5 женщин), голоса которых использовались для обучения синтеза речи. Для каждого диктора было записано по 9 парольных фраз (2—3 секунды речи). Примеры парольной фразы: „Город Екатеринбург, улица вокзальная, дом 22, вокзал“, „Заплатить три рубля и дать объявление в бюллетене“ и т.п. Важно отметить, что записанные фразы не использовались в дальнейшем при обучении системы синтеза речи. Итого было записано 63 фразы различных дикторов.

Влияние обучающих данных системы синтеза на надежность верификации. Цель экспериментов — установить зависимость ошибки ложного принятия верификации (FA) от длительности речевого фрагмента, используемого при обучении системы синтеза голоса. Для экспериментов была взята описанная ранее система верификации по голосу. Калибровка порогов срабатывания системы производилась на речевой базе YOHO [13], содержащей 138 дикторов (мужчины и женщины), каждый из которых произносил фиксированную парольную фразу вида „36-24-36“ длительностью около (1,5—2 секунды активной речи).

Были определены два порога системы верификации:

1) по равновероятной ошибке пропуска—отклонения (equal error rate, EER) — ThresholdEER. На калибровочной базе EER=4 %;

2) при вероятности ложного принятия не более 1 % — ThresholdFA1. Этот порог обычно используется, когда необходимо обеспечить максимальную защиту системы от доступа злоумышленника.

Далее для каждого диктора выполнялись попытки доступа в систему верификации с помощью синтезированных парольных фраз, подготовленных системой синтеза голоса; объем спонтанной речи, использованной для обучения, — от 1 минуты до 4 часов речи для каждого диктора. На рис. 2 приведены результаты эксперимента, где *1* — калибровочный FA, *2* — 1 мин, *3* — 3 мин, *4* — 8 мин, *5* — 30 мин, *6* — 4 часа (*N* — мера сходства).

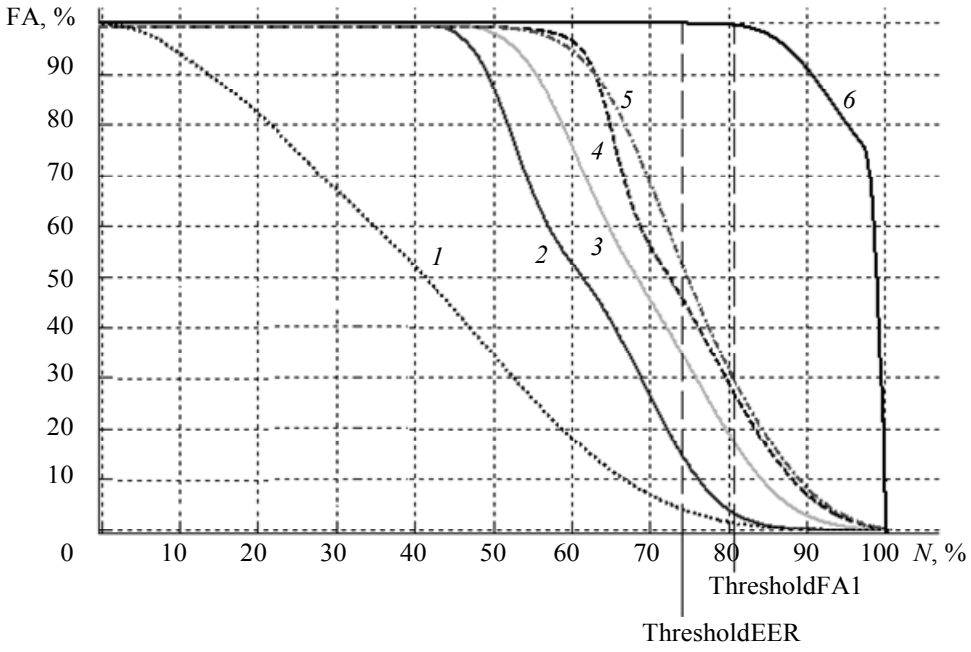

Рис. 2

В таблице представлены значения вероятности ложного принятия для двух порогов исследуемой системы верификации. Видно, что надежность системы верификации значительно снижается при использовании спонтанной речи длительностью от 8 минут и более; для системы верификации синтезированная речь практически перестает отличаться от живой речи человека при подготовке на данных большого объема (4 часа речи).

Объем речи для обучения синтеза	Ошибка FA (%) для порога	
	ThresholdEER	ThresholdFA1
1 минута	12,7 (8)	1,5 (1)
3 минуты	34,9 (22)	7,9 (5)
8 минут	44,4 (28)	19,1 (12)
30 минут	55,6 (35)	23,8 (15)
4 часа	100 (63)	98,4 (62)

На основании полученных результатов можно сделать выводы о том, что предлагаемый метод спуфинга позволяет не только серьезно ослаблять надежность системы верификации, но и обходить такую дополнительную меру защиты, как детектор присутствия диктора. Если система верификации передает пользователю пароль в виде звукового сообщения, возможно использование системы распознавания речи для полной автоматизации процесса спуфинга. В отличие от спуфинга путем конвертации признаков речи [14—16], предложенный подход, при использовании совместно с системой распознавания речи, позволяет исключить участие человека из диалога с системой верификации.

Выводы. В статье проанализирована устойчивость современных методов верификации к спуфингу при помощи гибридной системы синтеза речи на основе технологий Unit Selection и HMM. Как показали эксперименты, уже при использовании 8 и более минут обучающего материала, возможно существенно снизить надежность системы верификации, а при увеличении обучающего материала до четырех часов система практически не отличает синтезированный звук от речи диктора.

Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

СПИСОК ЛИТЕРАТУРЫ

1. *Матвеев Ю. Н.* Технологии биометрической идентификации личности по голосу и другим модальностям // Вестн. МГТУ. „Приборостроение“. № 3 „Биометрические технологии“. 2012. С. 46—61.

2. The NIST Year 2012 Speaker Recognition Evaluation Plan [Electronic resource]: <http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf>.

3. *Wu Z., Kinnunen T., Chng E. S., Li H., Ambikairajah E.* A Study on spoofing attack in state-of-the-art speaker verification: the telephone speech case // Proc. of the APSIPA ASC 2012. Hollywood, USA, 2012. P. 1—5.

4. *Kenny P.* Bayesian speaker verification with heavy tailed priors // Proc. of the Odyssey Speaker and Language Recognition Workshop. Brno, Czech Republic, 2010.

5. *Simonchik K., Pekhovsky T., Shulipa A., Afanasyev A.* Supervized Mixture of PLDA Models for Cross-Channel Speaker Verification // Proc. of the 13th Annual Conf of the Intern. Speech Communication Association, Interspeech-2012. Portland, Oregon, USA, 2012. P. 1682—1685.

6. *Алейник С. В., Матвеев Ю. Н., Раев А. Н.* Метод оценки уровня клиппирования речевого сигнала // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 3 (79). С. 79—83.

7. *Pelecanos J., Sridharan S.* Feature warping for robust speaker verication // Proc. Speaker Odyssey.The Speaker Recognition Workshop. Crete, Greece, 2001. P. 243—248.

8. *Matveev Yu. N., Simonchik K. K.* The speaker identification system for the NIST SRE 2010 // The 20th Intern. Conf. on Computer Graphics and Vision, GraphiCon’2010. St. Petersburg, 2010. P. 315—319.

9. *Козлов А. В., Кудашев О. Ю., Матвеев Ю. Н., Пеховский Т. С., Симончик К. К., Шулипа А. К.* Система идентификации дикторов по голосу для конкурса NIST SRE 2012 // Тр. СПИИРАН. 2013. Т. 25, № 2. С. 350—370.

10. *Villalba J., Lleida E.* Speaker verification performance degradation against spoofing and tampering attacks // Proc. FALA 10 Workshop. 2010. P. 131—134.

11. *Chistikov P. G., Korolkov E. A., Talanov A.O.* Combining HMM and unit selection tech nologies to increase naturalness of synthesized speech // Proc. of the Annual Intern. Conf. “Dialog-2013”. P. 2—10.

12. *Chistikov P. G., Korolkov E. A.* Data-driven Speech Parameter Generation For Russian TexttoSpeech System. Computational Linguistics and Intellectual Technologies // Proc. of the Annual Intern. Conf. ”Dialogue”. 2012. Vol. 1, Is. 11. P. 103—111.

13. *Campbell J., Higgins A.* YOHO Speaker Verification [Electronic resource]: <http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC94S16>.

14. *Wu Z., Chng E. S., Li H.* Speaker verification system against two different voice conversion techniques in spoofing attacks: Technical report [Electronic resource]: <http://www3.ntu>.

15. *Kinnunen T., Wu Z.-Z., Lee K. A., Sedlak F., Chng E. S., Li H.* Vulnerability of Speaker Verification Systems Against Voice Conversion Spoofing Attacks: the Case of Telephone Speech // Proc. ICASSP. Kyoto, Japan, 2012. P. 4401–4404.

16. *Aylett M. P., Yamagishi J.* Combining statistical parameteric speech synthesis and unit-selection for automatic voice cloning. 2008.

Сведения об авторах

Вадим Леонидович Щемелинин — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: shchemelinin@speechpro.com

Константин Константинович Симончик — канд. техн. наук, доцент; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; ООО „ЦРТ“, Санкт-Петербург; руководитель отдела; E-mail: simonchik@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.