УДК 57.087.1

С. А. Новоселов, В. А. Сухмель, А. В. Шолохов, Т. С. Пеховский

ПРИМЕНЕНИЕ DTW-МЕТОДА ДЛЯ МУЛЬТИСЕССИОННОГО ОБУЧЕНИЯ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ В ЗАДАЧЕ ТЕКСТОЗАВИСИМОЙ ВЕРИФИКАЦИИ ДИКТОРА

Представлен метод обучения скрытых марковских моделей по нескольким вариантам произнесения парольной фразы с помощью алгоритма динамического временного выравнивания сигналов. Метод позволяет создавать точные статистические модели речевых сигналов и снижать вероятность возникновения ошибок верификации.

Ключевые слова: *текстозависимая верификация диктора, MFCC, HMM, GMM, DTW.*

Введение. В последнее время в области распознавания диктора по голосу преобладают исследования текстонезависимых методов [1—8]. Основу всех алгоритмов идентификации личности по голосу составляет GMM-UBM-подход, основанный на совместном использовании смесей гауссовых распределений (Gaussian Mixture Model, GMM) и универсальной фоновой модели (Universal Background Model, UBM) [9]. Определенного успеха в этой области удалось достичь за счет использования методов снижения размерностей и различных классификаторов для параметров средних (супервекторов) GMM-моделей дикторов [1—3, 6, 8]. Однако эффективность таких систем зависит от длительности речевых сигналов [4, 5, 7].

Одним из вариантов реализации систем верификации диктора по голосовому паролю может быть комбинирование дикторонезависимого распознавателя речи и текстонезависимого распознавателя диктора. Такой подход является языкозависимым и требует достаточного количества ресурсов. На наш взгляд, перспективен подход, при котором на этапе обучения системы верификации создается статистическая модель речевого сигнала, способная описывать последовательность звуков парольной фразы, а также особенности их произнесения конкретным диктором. Возможность создания такой скрытой марковской модели (Hidden Markov Model, HMM) подтверждается результатами работы [10]. В ней показано, что эмиссионные распределения состояний модели HMM-GMM парольной фразы могут быть аппроксимированы на основе адаптации статистической модели голоса диктора, которая, в свою очередь, обучается классическими методами текстонезависимой идентификации [1—9]. Верификационный тест при использовании такой генеративной модели удобно проводить путем оценки логарифма правдоподобия для тестового речевого сигнала с учетом наиболее вероятной последовательности состояний, т.е. с учетом пути Витерби [11].

В настоящей статье исследуется иерархическая система текстозависимой верификации диктора на различных базах речевых данных и разрабатывается метод обучения HMM-GMM-модели парольной фразы при наличии нескольких вариантов произнесения (сессией), на основе метода динамического программирования DTW (Dynamic Time Warping) [12].

Иерархическая система верификации диктора. В основе иерархического подхода к обучению статистических моделей лежит трехуровневая структура представления акустических признаков речевых сигналов. Подход предполагает последовательную адаптацию моделей по критерию максимума апостериорной вероятности (MAP-адаптация). Для описания парольной фразы используется скрытая марковская модель, что обусловлено возможностью „запоминания“ временной структуры речи с помощью введения множества состояний модели и матрицы вероятностей переходов между этими состояниями [10—14]. Чтобы учесть лингвистическую информацию, которая содержится в голосовом пароле диктора, статистические

модели состояний HMM парольной фразы предложено строить путем адаптации заранее обученной текстонезависимой GMM-модели диктора. Эмиссионные плотности распределения вероятностей марковской модели в данном случае аппроксимируются смесями гауссовых распределений. При таком обучении HMM сохраняются достоинства схемы GMM-UBM, которая успешно применяется в решении задач текстонезависимого распознавания диктора.

Общее описание. На рис. 1 схематически представлен иерархический подход к обучению HMM. Все узлы этой схемы являются смесями гауссовых распределений. Верхний уровень системы — это универсальная фоновая модель, которая статистически описывает общее акустическое пространство признаков различных голосов [9].

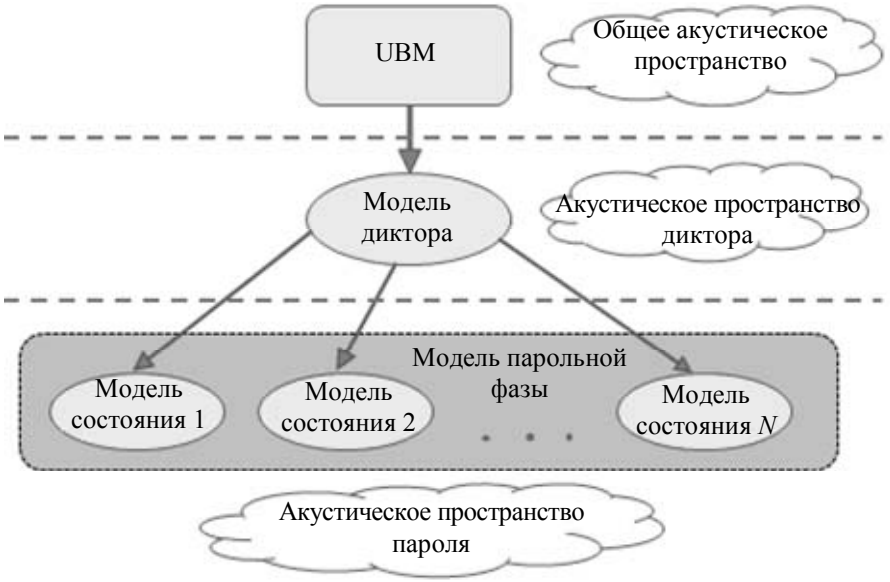

Рис. 1

Средний уровень системы предназначен для моделирования текстонезависимой информации о голосе конкретного человека. Параметры распределения акустических характеристик голоса могут быть оценены на основе критерия максимума апостериорной вероятности по известной универсальной фоновой модели. На этом уровне адаптируются только векторы математических ожиданий гауссиан, а все остальные параметры заимствуются у модели UBM. Модель голоса диктора используется далее на текстозависимом уровне для построения HMM-модели. Статистические модели состояний MM обучаются путем адаптации дикторской модели GMM, полученной на среднем уровне, к универсальной фоновой. При этом трансформируются только весовые коэффициенты гауссиан. Так формируется HMM-GMM-модель, которая описывает особенности голоса диктора и временную структуру последовательности звуков парольной фразы. Рассмотрим подробнее этапы обучения системы.

Обучение системы. Алгоритм обучения HMM-GMM состоит из трех этапов. На первом строится универсальная фоновая модель верхнего уровня иерархической системы. Расчет модели проходит на большом объеме речевой информации с помощью классического алгоритма максимизации ожидания (expectation maximization, EM) [1, 9].

Обучение дикторской модели текстонезависимого уровня осуществляется путем адаптации UBM-модели с использованием речевых сигналов конкретного диктора. На этом этапе необходимо учитывать решение детектора речевой активности для определения речевых сегментов сигналов, поскольку именно они необходимы для построения адекватной статистической модели голоса.

На последнем этапе обучается итоговая модель HMM (рис. 2). По мнению авторов работ [10, 12], на этом уровне необязательно исключать неречевые сегменты из парольной фразы, поскольку HMM способна моделировать паузы в речи. Весь речевой сигнал разбивается на

участки одинаковой длины. Каждое состояние HMM обучается методом адаптации дикторской модели верхнего уровня по данным соответствующего сегмента. MAP-адаптация выполняется только для весовых коэффициентов гауссиан дикторской модели. После того как модели состояний построены, HMM с равновероятной матрицей переходов оптимизируется с помощью классического алгоритма Витерби, а вероятности переходов между состояниями пересчитываются пропорционально относительной длине смежных сегментов.

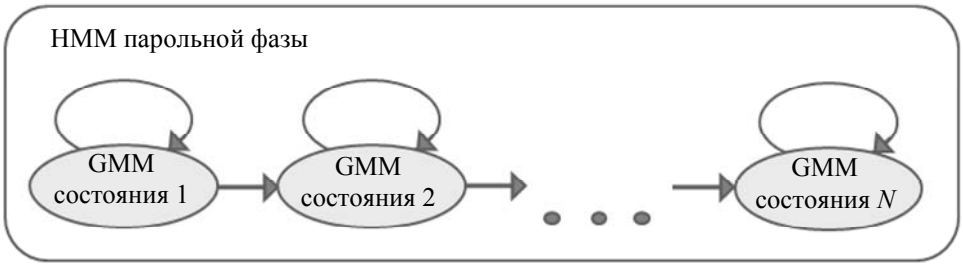

Рис. 2

Тестирование. Для определения метрик сравнения при проведении верификационных тестов сначала рассчитывается путь Витерби для тестовой последовательности речевых признаков. В декодировании Витерби участвуют все сегменты сигнала без исключения. Логарифм правдоподобия HMM λ вычисляется с учетом пути Витерби по состояниям и может быть записан как сумма логарифмов правдоподобия моделей состояний λ_S парольной фразы на речевых и неречевых кадрах:

$$\log p\left(X|\lambda\right) = \sum_{S\in\text{Speech}} \log p\left(X_S|\lambda_S\right) + \sum_{S\notin\text{Speech}} \log p(X_S \mid \lambda_S),$$

где X — параметры речевого сигнала, X_S — параметры, соответствующие состоянию S, Speech — множество состояний парольной фазы, содержащих речь.

Финальная метрика сравнения score формируется только из логарифмов правдоподобия на речевых сегментах путем нормализации с помощью логарифма правдоподобия универсальной фоновой модели λ_{UBM} и усреднения по всем таким состояниям S:

$$\text{score} = \frac{1}{S} \sum_{S\in\text{Speech}} \left[\log p\left(X_S|\lambda_S\right) - \log p(X_S \mid \lambda_{\text{UBM}})\right]$$

При наличии нескольких вариантов произнесения парольной фразы неясно, каким образом проводить обучение HMM-GMM-модели диктора, используя всю доступную речевую информацию. Для среднего уровня иерархической системы этот вопрос легко разрешается путем простой конкатенации информативных параметров нескольких произнесений парольной фразы. Таким образом, возможно получить более точно соответствующую дикторскую модель на текстонезависимом уровне [10]. Использование техники динамического временного выравнивания сигналов [15, 16] для обучения последнего уровня иерархии позволит принимать во внимание вариативность произнесения отдельных звуков парольной фразы и более точно формировать статистические модели состояний HMM с помощью критерия MAP по нескольким произнесениям.

Динамическое временное выравнивание сигналов. Целью DTW [16] является сравнение двух последовательностей $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ и $\mathbf{y} = \{y_1, y_2, ..., y_n\}$, где x_n и y_n — векторы временных параметров сигналов (например, мел-частотные кепстральные коэффициенты, MFCC [17]). Для того чтобы сравнить два вектора $\mathbf{x}$ и $\mathbf{y}$, введем понятие локальной дистанции:

$$d\left(\mathbf{x},\mathbf{y}\right) = \text{Norm}\left(\mathbf{x},\mathbf{y}\right) \to \left[0,+\infty\right).$$

С использованием локальной метрики сравнения векторов вычислим матрицу:

$$\mathbf{C}_{n,m} = d\left(\mathbf{x}_n, \mathbf{y}_m\right),$$

по которой определяется оптимальный путь сравнений последовательностей векторов **x** и **y**. Определив путь, можно проводить временную темпокоррекцию двух сигналов и определять участки речи, наиболее соответствующие друг другу по критерию выбранной метрики.

Обучение моделей состояний. Рассмотрим случай обучения GMM-моделей состояний при наличии двух вариантов произнесения парольной фразы с помощью DTW. Вначале речевые сигналы выравниваются по времени. Временная шкала первого сигнала, так же как и в базовом подходе [10], разбивается на равные сегменты. Данные для обучения статистических моделей состояний теперь будут формироваться по двум последовательностям векторов информативных параметров разных сигналов (рис. 3). Каждое состояние HMM обучается методом адаптации модели диктора по вновь сформированным данным соответствующего сегмента пароля.

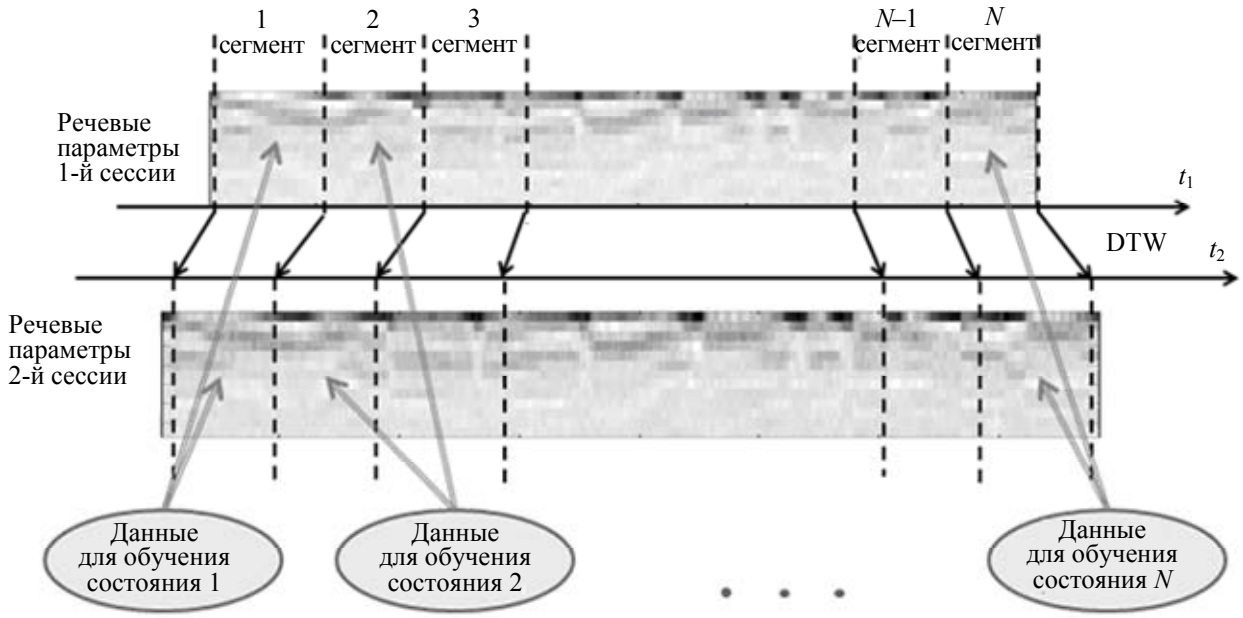

Рис 3

На этапе оптимизации моделей расчет пути Витерби производится только для первого речевого сигнала. В результате получается новое разбиение временной шкалы и происходит переобучение статистических моделей состояний. Формирование матрицы вероятностей переходов между состояниями осуществляется по правилам, описанным выше.

С помощью метода DTW аналогичным образом можно проводить обучение и для большего количества вариантов произнесения парольной фразы.

Параметры системы тестирования

Информативные параметры речевых сигналов, для их получения используются MFCC. Размерность векторов равна 13, предполагаются, что компоненты векторов некоррелированны между собой.

Детектор речевой активности построен на основе GMM-моделей.

Универсальная фоновая модель. Исходя из результатов статьи [10] размерность UBM выбрана равной 64. Для баз данных *ФИО_цифры_0_6* и *Цифры_0_9* обучалась на множестве 27 дикторов, для базы *POLYCOST* — строилась по рекомендациям работы [18].

Длина сегментов для обучения состояний HMM. Исследовалась зависимость равновероятной ошибки первого и второго рода (Equal Error Rate, EER) системы верификации дикторов

от длины сегментов, на которые разбивается парольная фраза в процессе обучения HMM (рис. 4). Модели обучались по базовому правилу на одном произнесении пароля. Тестирование проводилось по всем выбранным базам. Минимальное значение EER для трех случаев (*1 — Цифры_0_9*, *2 — ФИО_цифры_0_6*, *3 — POLYCOST*) достигается при сегментах порядка 0,03—0,04 с. Это означает, что на начальном этапе обучения HMM целесообразно разбивать речевые сигналы на одинаковые участки длительностью около 35 мс.

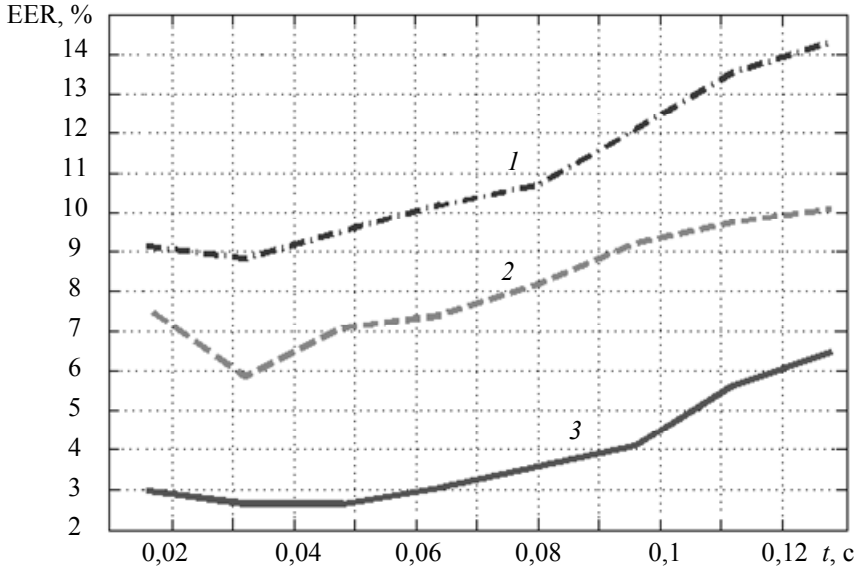

Рис. 4

Экспериментальные исследования выполнены на русскоязычных речевых базах *ФИО_цифры_0_6*, *Цифры_0_9* и англоязычной части базы *POLYCOST*. В настоящей работе эксперименты проводились на голосах мужских дикторов.

— Текстозависимая база *ФИО_цифры_0_6* содержит аудиозаписи 127 мужских голосов, которые произносят свои фамилию, имя, отчество и последовательность цифр от 0 до 6. База содержит 15 сессий каждого диктора. Интервал между записями сессий — не менее 1 дня и не более одной недели. Записи проводились в GSM-канале при использовании разных сотовых телефонов.

— База *Цифры_0_9* включает в себя записи 127 мужских дикторов (тех же, что в *ФИО_цифры_0_6*), которые произносят общую фразу, состоящую из последовательности цифр от 0 до 9. Условия сбора речевой базы те же, что и для *ФИО_цифры_0_6*.

— База *POLYCOST* [19] содержит 10 сессий, записанных 134 дикторами из 14 стран. Каждая сессия состоит из 14 произнесений; 4 повторений кода из 7 цифр, 5 последовательностей из 10 цифр, 2 фиксированных фраз, 1 международного номера телефона и 2 предложений с речью на родном языке диктора. Как и рекомендовано в [18], были исправлены ошибки базы и для проведения исследований тектозависимой верификации диктора выбрана парольная фраза “Joe took father's green shoe bench out”.

Результаты исследований EER рассматриваемой системы верификации приводятся для следующих вариантов обучения модели:

I — обучение дикторской модели и модели состояний по одному произнесению пароля;

II — обучение дикторской модели по двум произнесениям, моделей состояний — по одному из произнесений;

III — обучение дикторской модели и моделей состояний по двум произнесениям, с применением выравнивания MFCC-последовательностей по времени.

Из табл. 1 видно, что при обучении HMM-GMM-моделей по нескольким вариантам произнесения парольной фразы применение DTW-метода позволяет снизить уровень равновероятной ошибки верификации. Видно также, что при обучении по двум произнесениям только

дикторской модели второго уровня EER системы верификации в некоторых случаях увеличивается по сравнению с EER системы, обученной на одном произнесении. Это может быть следствием получения менее адекватных моделей состояний HMM при таком варианте обучения.

Таблица 1

Сравнение EER (%) для разных вариантов обучения

База	I	II	III		
			L_1	L_2	cos
ФИО_цифры_0_6	5,85	6,61	5,16	4,09	4,25
Цифры_0_9	8,81	9,08	7,40	7,30	7,53
POLYCOST	2,63	2,58	2,63	1,76	1,69

Дополнительно исследовалось влияние выбора локальной метрики сравнения DTW-методе на ошибки системы верификации. Выявлено, что норма схожести MFCC-векторов является лучшей по сравнению с нормой L_2 и косинусной метрикой [20].

При сравнении значений EER, полученных на базах *ФИО_цифры_0_6* и *Цифры_0_9*, необходимо учесть, что при работе с первой моделируется случай, когда речевой пароль не известен злоумышленнику, а вторая — когда злоумышленник знает парольную фразу. Голоса дикторов и условия записи аудиофайлов в этих базах совпадают. Из результатов табл. 1 видно, что значение EER в случае неизвестного пароля в среднем на 3 % меньше, чем при известном пароле.

Результаты, полученные на базе *POLYCOST*, лучше, поскольку она меньше по объему и речевые сигналы в ней менее искажены каналом передачи.

В табл. 2 приведены значения EER, полученные при тестировании системы верификации обученной на пяти вариантах произнесения парольной фразы. В методе DTW для сравнения векторов признаков использовалась норма L_2. Ошибки верификации снижаются при увеличении объема обучающих данных. Необходимо отметить, что для текстозависимой базы *ФИО_цифры_0_6* при обучении моделей на четырех вариантах произнесения пароля наблюдается снижение EER на 40 % по сравнению со случаем обучения на одном варианте произнесения. При увеличении числа произнесений до 5 относительное изменение EER составляет ≈47 %. Для баз *Цифры_0_9* и *POLYCOST* при использовании для обучения 5 сессий вместо одной удается снизить ошибку EER на 23 и 50 % соответственно.

Таблица 2

Сравнение EER (%) при обучении на пяти вариантах произнесения

База	1	2	3	4	5
ФИО_цифры_0_6	5,85	4,09	3,92	3,50	3,04
Цифры_0_9	8,81	7,30	7,29	6,92	6,77
POLYCOST	2,63	1,76	1,70	1,32	1,31

Выводы. В работе исследована иерархическая система текстозависимой верификации диктора на трех различных речевых базах. Найдена оптимальная по критерию равновероятной ошибки системы верификации длина сегментов, на которые необходимо разбивать речевой сигнал при обучении HMM-GMM-моделей. Представлен новый метод обучения HMM-GMM-модели парольной фразы при наличии нескольких вариантов произнесения с помощью временной темпокоррекции речевых сигналов. Показано, что в качестве локальной метрики сравнения MFCC-векторов в методе DTW эффективно использовать норму L_2. Анализ результатов показал, что для тектозависимого случая при использовании четырех вариантов произнесения пароля на этапе обучения HMM-GMM вместо одного удается снизить EER системы верификации на 30 %.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Kenny P., Boulianne G., Ouellet P., Dumouchel P.* Speaker and Session Variability in GMM-Based Speaker Verification // IEEE Transact. on Audio, Speech, and Language Processing. 2007. Vol. 15, N 4. P. 1448—1460.

2. *Vogt R. J., Lustri C. J., Sridharan S.* Factor Analysis Modelling for Speaker Verification with Short Utterances // Proc. Of Speaker and Language Recognition Workshop "Odyssey-2008". Stellenbosch, South Africa, 2008. P. 1—5.

3. *Матвеев Ю. Н.* Технологии биометрической идентификации личности по голосу и другим модальностям // Вестн. МГТУ им. Н. Э. Баумана. Сер. „Приборостроение". 2012. № 3(3). С. 46—61.

4. *Vogt R., Sridharan S., Mason M.* Making confident speaker verification decisions with minimal speech // IEEE Transact. On Audio Speech, and Language Processing. 2010. Vol. 18, N 6. P. 1182—1192.

5. *McLaren M., Vogt R., Baker B., Sridharan S.* Experiments in SVM-based Speaker Verification Using Short Utterances // Proc. of Speaker and Language Recognition Workshop (Odyssey—2010). Brno, Czech Republic, 2010. P. 83—90.

6. *Kanagasundaram A., Vogt R., Dean D. B., Sridharan S., Mason M. W.* I-vector based speaker recognition on short utterances // Proc. of 12th Annual Conf. of International Speech Communication Association (INTERSPEECH 2011). Firenze Fiera, Florence, 2011. P. 2341—2344.

7. *Kanagasundaram A., Vogt R. J., Dean D. B., Sridharan S.* PLDA based speaker recognition on short utterances // Proc. of Speaker and Language Recognition Workshop "Odyssey-2012". Singapore, 2012. P. 28—33.

8. *Stafylakis T., Kenny P., Senoussaoui M., Dumouchel P.* PLDA using Gaussian Restricted Boltzmann Machines with application to Speaker Verification // Proc. Of 13th Annual Conf. of Intern. Speech Communication Association (INTERSPEECH 2012). Portland, Oregon, USA. 2012. P. 2341—2344.

9. *Reynolds D., Quatieri T., Dunn R.* Speaker Verification using Adapted Gaussian Mixture Models // Digital Signal Processing. 2000. Vol. 10. P. 19—41,

10. *Larcher A. O., Bonastre J.-F., Mason J. S. D.* From GMM to HMM for embedded password-based speaker recognition // Proc. 16th Europ. Signal Processing Conf. (EUSIPCO-2008). Lausanne, Switzerland, 2008. P. 1—5.

11. *Juang B. H., Rabiner L. R.* Hidden Markov Models for Speech Recognition // Technometrics. 1991. Vol. 33, N 3. P. 251—272.

12. *Geppener V. V., Simonchik K. K., Haidar A. S.* Design of speaker verification systems with the use of an algorithm of Dynamic Time Warping (DTW) // Pattern Recognition and Image Analysis. 2007. Vol. 17, N 4. P. 470—479.

13. *Larcher A. O., Bonastre J.-F., Mason J. S. D.* Reinforced Temporal Structure Information for Embedded Utterance-Based Speaker Recognition // Proc. of 9th Annual Conf. of Intern. Speech Communication Association (INTERSPEECH 2008). Brisbane, Australia, 2008. P. 371—374.

14. *Subramanya A., Zhengyou Z., Surendran A. C., Nguyen P., Narasimhan M., Acero A.* A Generative-Discriminative Framework using Ensemble Methods for Text-Dependent Speaker Verification // Proc. of the Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP—2007). Honolulu, Hawaii, USA, 2007. Vol. 4. P. 225—228.

15. *Винцюк Т. К.* Распознавание слов устной речи методами динамического программирования // Кибернетика. 1968. № 1. С. 81—88.

16. *Muller M.* Information Retrieval for Music and Motion. Springer, 2007. 318 p.

17. *Матвеев Ю. Н.* Исследование информативности признаков речи для систем автоматической идентификации дикторов // Изв. вузов. Приборостроение. 2013. Т. 56, № 2. С. 47—51.

18. *Melin H., Lindberg J.* Guidelines for experiments on the POLYCOST База. KTH/Centre for Speech Technology [Электронный ресурс]: <http://www.speech.kth.se/cost250/polycost/be/v2.0/>.

19. *Petrovska D., Hennebert J., Melin H., Genoud D.* Polycost: A Telephone-Speech Database for Speaker Recognition // Proc. Workshop on Speaker Recognition and its Commercial and Forensic Applications. Avignon, France, 1998. P. 211—214.

Сведения об авторах

Сергей Александрович Новосёлов — канд. техн. наук; ООО „ЦРТ-инновации", Санкт-Петербург; научный сотрудник; E-mail: novoselov@speechpro.com

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-