Разработка прототипа системы распознавания личности по голосу

Людмила Сухостат

Институт Информационных Технологий НАН Азербайджана

lsuhostat@hotmail.com

***Аннотация*— В работе рассматривается разработка прототипа системы распознавания личности по голосу основанная на кепстральных коэффициентах по шкале мел (Mel-frequency Cepstral Coefficients, MFCC) и смешанных гауссовых моделях (Gaussian Mixture Models, GMM). Приводится описание созданной речевой базы данных для азербайджанского языка. Эксперименты показали достаточно высокие результаты работы системы.**

Ключевые слова*— *система распознавания личности по голосу; кепстральные коэффициенты по шкале мел; гауссовы смешанные модели; речевая база данных.

I. ВВЕДЕНИЕ

Системы распознавания личности по голосу широко применяются на практике [1]: доступ к базам данных, к банковским счетам, в криминалистической экспертизе.

Автоматическое распознавание диктора – вычислительная задача проверки заявленной подлинности пользователя, использующая характеристики, извлекаемые из его голоса.

При автоматическом распознавании диктора речевой сигнал обрабатывается, чтобы извлечь характерную информацию о говорящем [2,3]. Эта информация используется для генерирования идентификатора диктора, который не может быть воспроизведен любым источником, кроме оригинала. Это делает процесс распознавания диктора безопасным способом аутентификации пользователей в отличие от паролей или токенов, голос не может быть украден, дублирован или забыт.

В отличие от других биометрических технологий, которые в основном основаны на изображении и требуют дорогостоящего оборудования, таких как датчик отпечатков пальцев или сканер радужной оболочки глаз, системы распознавания диктора предназначены для использования практически любых стандартных телефонов или телефонных сетей общего пользования. Умение работать со стандартным телефонным оборудованием позволяет поддерживать широкий спектр биометрических голосовых приложений в самых разных условиях.

В Институте Информационных Технологий Национальной Академии Наук Азербайджана на протяжении нескольких лет проводятся исследования в области распознавания личности по голосу [4-6]. В результате накопленного опыта был разработан прототип системы распознавания диктора, и собрана речевая база данных для азербайджанского языка.

Целью исследований является создание простой и удобной автоматической системы распознавания диктора.

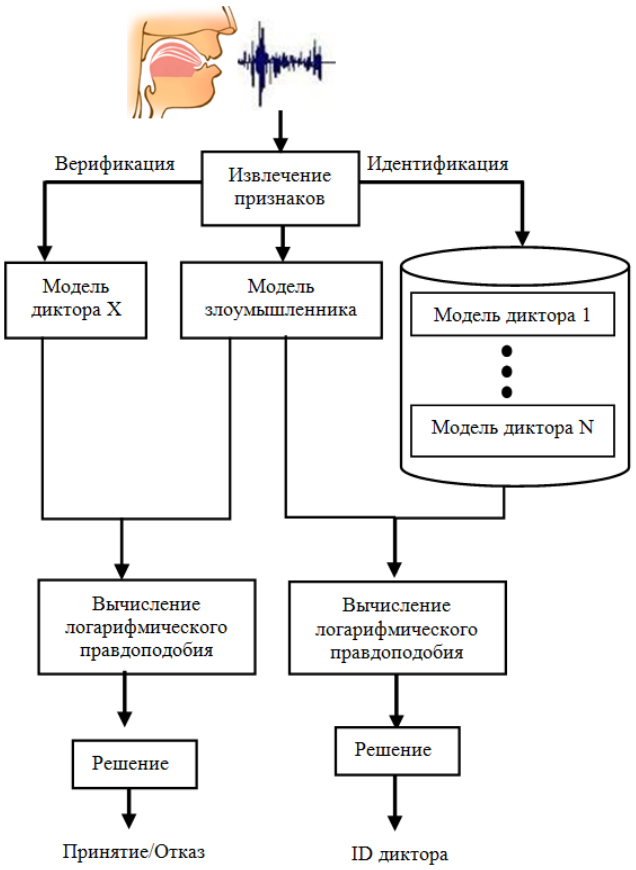

Рис 1. Общая архитектура системы верификации и идентификации диктора

В данной статье дается общая архитектура работы системы распознавания диктора, приводится описание созданной речевой базы данных, краткое описание методов MFCC и GMM.

II. АРХИТЕКТУРА СИСТЕМЫ РАСПОЗНАВАНИЯ ДИКТОРА

Общая схема разработанной системы представлена на рис. 1. Приложение было создано в среде Visual Studio 2010 на языке программирования C++.

Система состоит из нескольких основных подсистем: регистрации, верификации и идентификации.

Идентификационные данные и голосовые модели пользователей хранятся в базе данных SQLite.

В подсистеме регистрации происходит запись речевых образцов и создание речевой модели после обработки этих образцов. Идентификационные данные нового пользователя вместе с голосовой моделью добавляются в базу данных. В подсистеме также генерируется модель злоумышленника на основе универсальной фоновой модели (Universal Background Model, UBM).

В подсистеме идентификации на основе речевого образца проводится поиск наиболее близких в некоторой метрике пользователей. Результатом может быть список пользователей или пустой список. Список сортируется по возрастанию расстояний схожести от искомого пользователя.

В подсистеме верификации проверяется подлинность говорящего. Для этого диктор должен ввести в текстовое поле номер своего идентификатора и произнести ключевое слово. В случае успешного определения выводятся идентификационные данные пользователя и ему предоставляется право доступа.

Также в этой системе возможно обновление голосовых моделей пользователя на основе новых записей образцов речи.

III. СОЗДАНИЕ РЕЧЕВОЙ БАЗЫ ДАННЫХ ДЛЯ АЗЕРБАЙДЖАНСКОГО ЯЗЫКА

Производительность автоматической системы верификации очень сильно зависит от речевой базы данных. Есть много факторов, которые влияют на производительность системы автоматического распознавания. К ним относятся условия записи, окружающая среда, устройства записи, длительность, пол диктора, возрастная группа и т.д. Не зная условий записи, бессмысленно ожидать хорошего результата системы автоматического верификации.

Некоторые из важных причин, почему необходим речевой корпус, приведены ниже:

а) корпус с помощью голосовых аудиоданных может отражать язык контента, который соответствует географическому окружению;

б) практические характеристики связной речи, которые не могут быть отражены в текстовой базе данных, хорошо отражают индивидуальные характеристики пользователей;

в) в отличие от текстовых корпусов, речевой корпус отражает просодическую информацию, а также указывает на предпочтительный стиль произнесения выбранной социально-культурной модели.

Описание базы данных

Речевой корпус для азербайджанского языка был собран Институтом Информационных Технологий НАНА. Содержит записи 86 дикторов (21 мужчина и 65 женщин). Запись производилась с помощью программного обеспечения Cool Edit Pro в офисных условиях. Все данные включают: изолированные цифры, изолированные слова, комбинации цифр и текстовый фрагмент.

База данных была записана в формате ".wav" с помощью микрофона с частотой дискретизации 11025 Гц при разрешении 16 бит за одну сессию. Все дикторы являются носителями языка. Общий объем данных для каждого диктора составляет приблизительно 2500 Кб. Средняя продолжительность речи для каждого говорящего составляет около 100 сек. Имеются 86 дикторов, которым присвоены идентификаторы от 1001 до 1086.

IV. ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ РЕЧЕВЫХ СИГНАЛОВ

Исследования показали, что человеческое восприятие звуков речевых сигналов не соответствует линейной шкале. Поэтому применяя MFCC коэффициенты [7,8] можно более точно аппроксимировать слуховую систему человека. Это позволяет лучше обработать данные.

Речевой сигнал состоит из звуков с различными частотами. Для каждого звука с фактической частотой f, измеряемой в Гц, субъективная высота измеряется по шкале, называемой «мел». Это линейная частота, лежащая ниже 1000 Гц, а логарифмически расположенная выше 1000 Гц. Для вычисления мелов для частоты f используется следующая формула аппроксимации

$$mel(f)=2595*\log_{10}(1+\frac{f}{700}). \quad (1)$$

Для извлечения MFCC коэффициентов сигнал разбивается на фреймы, к которым применяется оконное сглаживание, чтобы минимизировать спектральные искажения. В результате получаем сигнал вида

$$y(n)=x(n)w(n),\ 0\le n\le N-1. \quad (2)$$

Применяя окно Хэмминга

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right),0\le n\le N-1. \quad (3)$$

окончательно с помощью дискретного косинусного преобразования (Discrete Cosine Transform, DCT) осуществляется перевод логарифмического спектра обратно во временную область, и вычисляются MFCC коэффициенты

$$C_i=\sqrt{\frac{2}{N}}\sum_{j=1}^{L}\log(m_j)\cos\left(\frac{nj}{L}(i-0.5)\right). \quad (4).$$

V. ГАУССОВЫ СМЕШАННЫЕ МОДЕЛИ

Для моделирования распределения векторов признаков, полученных от каждого пользователя, в работе применяются смешанные гауссовы модели (Gaussian

Mixture Models, GMM) [9,10]. GMM можно рассматривать как непараметрическую многомерную функцию плотности вероятности (Probability Density Function, PDF), которая способна моделировать произвольные распределения и является наиболее предпочтительным методом моделирования дикторов.

Одним из основных преимуществ GMM является способность к созданию гладких аппроксимаций произвольных форм распределений. GMM по сравнению с другими подходами имеют быструю фазу обучения, модели могут быть легко масштабированы и обновлены при добавлении новых дикторов [11].

GMM распределение векторов признаков для диктора S представляет собой взвешенную линейную комбинацию *M* унимодальных плотностей гауссиан $b_i^S(x)$, каждая из которых параметризируется вектором математических ожиданий μ_i^S и ковариационной матрицей $\sum_i^S$. Эти параметры вместе представлены следующей записью:

$$\lambda_S = \{p_i^S, \mu_i^S, \sum_i^S\}, \quad i = 1,...,M \tag{5}$$

где p_i^S – смешанные веса, удовлетворяющие условию $\sum_{i=1}^{M} p_i^S = 1$. У каждого диктора своя модель λ_S.

Для вектора признаков x смешанная плотность для диктора S вычисляется как

$$p(x|\lambda_S) = \sum_{i=1}^{M} p_i^S b_i^S(x), \tag{6}$$

где

$$b_i^S(x) = \frac{1}{(2\pi)^{D/2}\left|\sum_i^S\right|^{1/2}} \exp\left\{\frac{1}{2}(x-\mu_i^S)'\sum_i^S(x-\mu_i^S)\right\}. \tag{7}$$

Для данной последовательности векторов признаков $X = \{x_1, x_2, \ldots, x_T\}$, которые предполагаются независимыми, логарифмическое правдоподобие [12] модели диктора λ_S представлено в виде

$$L_S(X) = \log p(X|\lambda_S) = \frac{1}{T}\sum_{t=1}^{T} \log p(x_t|\lambda_S). \tag{8}$$

Для идентификации диктора последнее уравнение вычисляется для каждого диктора зарегистрированного в системе. В данной работе используется GMM с 32 смесями для каждой модели.

Идентичность диктора определяется моделью с наибольшим значением. Для нахождения максимального правдоподобия моделей применяются различные алгоритмы. Одним из них является алгоритм EM (Expectation-Maximization) [12].

При этом для каждого диктора S находим следующие значения:

смешанные веса:

$$p_i = \frac{1}{T}\sum_{t=1}^{T} pr(i|x_t, \lambda), \tag{9}$$

математические ожидания:

$$\mu_i = \frac{\sum_{t=1}^{T} pr(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} pr(i|x_t, \lambda)}, \tag{10}$$

ковариационная матрица:

$$\Sigma_i = \frac{\sum_{t=1}^{T} pr(i|x_t, \lambda) x_i^2}{\sum_{t=1}^{T} pr(i|x_t, \lambda)} - \mu_i^2, \tag{11}$$

где апостериорная вероятность для компонента i имеет вид

$$pr(i|x_t, \lambda) = \frac{p_i b_i(x)}{\sum_{k=1}^{M} p_k b_k(x)}. \tag{12}$$

В современных системах распознавания диктора для моделирования альтернативной гипотезы применяется UBM. Она содержит в себе дикторо- и канально-независимые характеристики. UBM – это набор GMM (1024 гауссиан), обученных на дикторо-независимых распределениях признаков.

VI. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Для проведения экспериментов из рассматриваемой речевой базы данных были взяты речевые образцы с названиями городов.

Вначале каждый речевой сигнал предварительно обрабатывался. Длина окна Хэмминга составила 25 мсек, перекрытия – 12,5 мсек. Далее извлекались 20 MFCC коэффициентов.

Обучение проходило на 5 образцах для каждого диктора. Длина обучающих речевых данных составила приблизительно 10 сек. Тестирование проводилось на образцах длительностью 3 сек.

ТАБЛИЦА I. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Число гауссиан = 32					
Длительность обучения	***Длительность тестирования***	***Точность распознавания***	***FAR***	***FRR***	***EER***
10 сек	3 сек	93,7%	8,05%	7,72%	7,89%

Результаты экспериментов показаны в таблице I. Приводятся точность распознавания, вероятности ложного допуска (False Accept Rate, FAR) и ложного отказа (False Reject Rate, FRR), а также равная вероятность ошибок первого и второго рода (Equal Error Rate, EER) [13].

ЗАКЛЮЧЕНИЕ

В работе приводится описание разработанного прототипа системы распознавания личности по голосу. Рассматриваются результаты тестирования разработанной системы на собранной речевой базе данных для азербайджанского языка. Эксперименты показали достаточно высокие результаты ее работы.

В дальнейшем будут продолжаться исследования в направлении разработки автоматической системы распознавания диктора с целью дополнительного повышения уровня ее безопасности и улучшения показателей распознавания.

БИБЛИОГРАФИЯ

[1] J.P. Campbell, "Speaker recognition: A tutorial," Proc. IEEE, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[2] J. Benesty, M. Sondhi, Y. Huang, Springer handbook of speech processing, Springer, 2007. 1176 p.

[3] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, New Jersey, 1993. 277 p.

[4] Я.Н. Имамвердиев, Л.В. Сухостат, "Речевые базы данных для систем распознавания диктора", Вопросы защиты информации, no. 4, 2011, с. 27 – 32.

[5] Л.В. Сухостат, "Разработка методов и алгоритмов для синтеза систем биометрической идентификации личности по голосу", Науч. семинар, 30 ноября 2012, Баку, с. 29-30.

[6] Я.Н. Имамвердиев, Л.В. Сухостат, "Об одном методе извлечения признаков для систем распознавания диктора", İnformasiya texnologiyaları problemləri, no. 2, pp. 14-19, 2012.

[7] S. Furui, Digital speech processing, synthesis, and recognition. MarcelDekker, 2000. 452 p.

[8] S. Furui, "Cepstral analysis techniques for automatic speaker verification", IEEE Tran. acoust., speech, signal processing, vol.27, pp. 254-277, 1981.

[9] D.A. Reynolds "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication. vol 17, pp. 91-108, 1995.

[10] D.A. Reynolds "A Gaussian mixture modeling approach to text-independent speaker identification", Ph.D. Thesis. – Georgia Institute of Technology, September, 1992.

[11] A. Fazel and S. Chakrabartty, "An overview of Statistical Pattern Recognition Techniques for Speaker Verification", IEEE Circuits and System Magazine, pp. 62-81, 2011.

[12] A.P. Dempster, N.M. Laird, D.B. Rubin "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, vol. 39, pp. 1-38,1977.

[13] C.M. Bishop, Pattern Recognition and Machine learning, Springer Science, New York, 2006.