УДК 537.531

УНИВЕРСАЛЬНАЯ ФОНОВАЯ МОДЕЛЬ ДЛЯ ЗАДАЧ ВЕРИФИКАЦИИ ДИКТОРА

Крищенович В.А., магистрант

Белорусский государственный университет информатики и радиоэлектроники

г. Минск, Республика Беларусь

Захарьев В.А. – канд. техн. наук, доцент

Аннотация. Доклад посвящен роли универсальной фоновой модели в системах голосовой верификации диктора.

Ключевые слова. Голосовая верификация, голосовые признаки, универсальная фоновая модель, обучение УФМ.

Область речевых технологий на протяжении многих лет остается одной из самых актуальных направлений исследований. Это связано с широким кругом ее применимости в жизнедеятельности человека. Еще ни один компьютер не способен так быстро распознавать звуки, как это может делать человеческое ухо. И даже этот один факт говорит о том на сколько существующие системы не совершенны и требуют доработок.

Одной из наиболее сложных задач в области обработки речи является верификация диктора. Голос уступает по надежности идентифицирующего признака отпечаткам пальцев или сетчатки глаза. Но в качестве дополнительного фактора верификации в реальных система разграничения доступа, голосовая верификация дает ощутимые преимущества: возможность выполнения процедуры верификации без непосредственного визуального контакта с системой, отсутствие необходимости тактильного взаимодействия с системой – свободные руки, и пр.

Современные системы голосовой идентификации и верификации работают в двух режимах: обучение и рабочий режим.

В режиме обучения выделяются характерные признаки голоса диктора. Далее на основе этих признаков формируется его голосовая модель (голосовой отпечаток) и затем данная модель сохраняется в базе данных.

Кроме того, в режиме обучения на основе выделенных признаков голоса диктора также составляется так называемая универсальная фоновая модель -- УФМ (Universal Background Model -- UBM), которая описывает некоторые усредненные голосовые характеристики всех дикторов, находящихся в базе[2].

В рабочем режиме выделяются характерные признаки голосового сигнала диктора, затем выполняется проверка принадлежности признаков к конкретной заданной голосовой модели - верификация диктора. Также в этом режиме на основании универсальной фоновой модели проводится вычисление степени уникальности голосового сигнала. Данная степень уникальности позволяет судить о достоверности верификации и является частью аппарата принятия конечного решения.

Функциональная схемы работы такой системы представлена на рисунке 1.

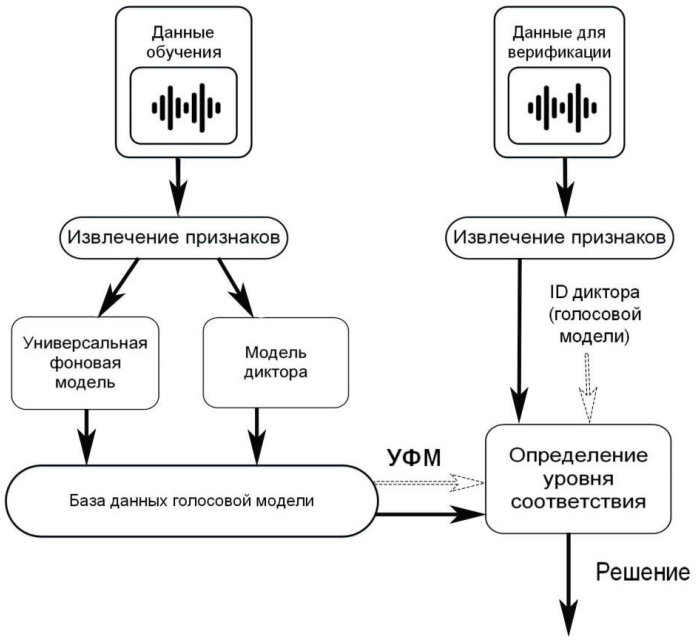

Рисунок 1 – Функциональная схема системы голосовой верификации диктора

При извлечении признаков голоса диктора обычно производится обработка сигнала: разбиение всего сигнала на отрезки и другие действия, в зависимости от типа извлекаемых признаков. Эти признаки используются для обучения универсальной фоновой модели. После обучения универсальной фоновой модели производится обучение моделей дикторов, зарегистрированных в системе, с помощью адаптации от универсальной фоновой модели. Для тестирования модели вычисляется логарифм отношения правдоподобия между заданной моделью диктора и универсальной фоновой моделью для заданного речевого сигнала [2, 4].

Также при создании универсальной фоновой модели необходимо помнить, что данные, используемые для обучения модели, должны быть сбалансированными по отношению к дальнейшему применению системы: сбалансированы по гендерному типу, оборудованию и стандартным условиям. Т.е. длительность обучающей выборки для дикторов мужчин и дикторов женщин должна быть примерно одинаковой для системы голосовой верификации [2]. Аналогичным образом, обучающие данные должны быть сбалансированы и по типу используемых при записи дикторов микрофонов [2].

Для обучения универсальной фоновой модели используется речевой корпус, который содержит аудиозаписи большого количества дикторов. Т.к. универсальная фоновая модель – это большая модель Гауссовой смеси, обученная для представления дикторонезависимого распределения признаков, то системы использующие данные модели, называются системами верификации диктора на основе модели Гауссовых смесей и универсальной фоновой модели (GMM-UBM) [1, 3].

Обучение универсальной фоновой модели возможно методом объединения результатов обучения отдельных моделей для разных выборок в одну. Например, объединение отдельных модели, обученные на выборках с дикторами мужчинами и дикторами женщинами, или обучение отдельных моделей для записей на различные типы микрофонов [2]. Также известны другие подходы, связанные с обучением моделей для групп дикторов [3, 4].

Простое обучение универсальной фоновой модели возможно на всей обучающей выборке с помощью метода максимального правдоподобия - EM (Expectation-Maximization) алгоритма [5].

Задача метода состоит в нахождении по заданным обучающим данным таких параметров модели, при которых функция правдоподобия модели достигает максимума.

На первом шаге алгоритма вычисляется ожидаемое значение функции правдоподобия. На втором шаге вычисляется оценка максимального правдоподобия для каждой компоненты модели. Затем вычисляются новые параметры модели, которые используются на первом шаге следующей итерации алгоритма [5].

В GMM-UBM системе для создания модели диктора производится адаптация параметров универсальной фоновой модели на обучающих данных конкретного диктора. Данная адаптация

известна также как Байесово обучение или оценка апостериорного максимума. Это позволяет не только увеличить точность распознавания диктора по сравнению с неадаптированными моделями, но и ускорить оценку соответствия моделей [7].

Также, как и в EM-алгоритме, адаптация состоит из двух шагов, однако на втором шаге вычисленные параметры смешиваются с исходными параметрами, взятыми из универсальной фоновой модели, по определенному коэффициенту[5].

Для существующих систем верификации используются базы речевых данных в несколько сотен часов, при этом обучение универсальной фоновой модели даже на современном оборудовании может длиться ни одну неделю [6]. Также погрешность вычислений влияет на качество записанных голосовых данных.

Таким образов, в настоящее время существует потребность в разработке качественного метода обучения универсальной фоновой модели на основе GMM-UBM модели с более высокой скоростью вычислений и меньшей погрешностью, способной работать с голосовыми материалами среднего качества (например, запись телефонного разговора) и менее чувствительной к изменению условий регистрации голосового сигнала, что повлияет на улучшение надежности системы.

Список использованных источников:

1. Bimbot F. et al. A tutorial on text-independent speaker verification // EURASIP J. on Applied Signal Processing. – 2004. – No. 4. – P. 430–451.

2. Reynolds D., Quatieri T.F., Dunn R.B., Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Process. 10 (2000), pp 19-41.

3. Reynolds D., Rose R. Robust text-independent speaker identification using Gaussian mixture speaker models // IEEE Trans. On Speech and Audio Processing. – 1995. – No. 3. – P. 72–83.

4. Rosenberg A. E., Parthasarathy S. Speaker background models for connected digit password speaker verification // Acoustics, Speech, and Signal Processing (ICASSP-96), IEEE International Conference on. – 1996. – Т. 1. – С. 81-84.

5. Sadjadi S.O., Slaney M., Heck L. MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research // Speech and Language Processing Technical Committee Newsletter. – 2013. – Т. 1. – № 4. – С. 1–32.

6. Габдуллин, В.В. Применение технологии CUDA для задач голосовой биометрии на примере построения универсальной фоновой модели диктора / В.В. Габдуллин, А.И. Капустин, А.И. Королев // Параллельные вычислительные технологии (ПаВТ'2011). – СпБ., 2011. – С. 107-116.

7. Рохманенко, И.В. Алгоритмы и программные средства верификации диктора по произвольной фразе / И.В. Рахманенко; науч. рук.Р.В. Мещеряков; ТУСУР. – Томск, 2017. – 111 с.