САНКТ-ПЕТЕРБУРГСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Математико-Механический факультет
Кафедра Системного Программирования

Алгоритм верификации диктора для встроенных систем

Курсовая работа студента 344 группы
Абрамова Ивана Александровича

Научный руководитель:
аспирант Королев А. И.

Санкт-Петербург
2014 г.

Содержание

1 Введение

В современном мире достаточно остро стоит вопрос защиты информации. Компании желают обезопасить себя от несанкционированного доступа к данным. Поэтому для аутентификации пользователя в системе хочется использовать не только символьный пароль, но и биометрические данные, такие как голос, отпечатки пальцев, сетчатку глаза. Использование голосовых данных мотивировано тем, что нет необходиости оснащать защищаемую систему дополнительным оборудованием. У большинства устройств уже имеется встроенный микрофон, с которого можно записывать речевые данные. Представленный подход реализован компанией Nuance[1] и успешно применяется для подтверждения транзакций в банках или для получения доступа к информации в сфере здравоохранения.

Не смотря на то, что идея внедрения систем защиты, использующих голосовую информацию набирает популярность с каждым годом, готовых решений, предоставляющих требуемую функциональность для встроенных систем, ещё нет. Во многом это связано с тем, что традиционные методы верификации диктора не дают приемлемую точность распознавания для задач с ограниченным количеством входных данных[1]. Использование существующих библиотек для распознавания диктора, таких как [2, 3], осложнено из-за того, что данные средства не учитывают архитектурные особенности встроенных систем.

1.1 Область исследований

Задача верификации диктора заключается в том, чтобы определить по звуковому сигналу, является ли говорящий тем, за кого он себя выдает, или нет. Для решения поставленной задачи существует несколько основных подходов:

- Текстонезависимый подход
- Тектозависимый подход

Текстонезависимый подход применяется в случаях, когда система верификации не обладает информацией о том, какую фразу должен произнести диктор. Такой подход чувствителен к количеству речевого материала, необходимого при обработке. Эффективность данного решения заметно ухудшается при использовании записей длительностью менее 30-ти секунд[1].

Текстозависимый подход, напротив, основывается на начальном знании фразы-пароля. Данный подход часто используется при распознавании речи и моделировании отдельных предложений и слов. Текстозависимый подход предоставляет информацию о временной структуре фразы, используя которую можно улучшить эффективность текстонезависимого решения[4].

[1] Nuance Communications Corporation - ведущий разработчик и поставщик речевых технологий. http://www.nuance.com (дата обращения: 15.03.2014).

1.2 Постановка задачи

В рамках данной работы были поставлены следующие задачи:

- Изучить существующие методики верификации диктора
- Реализовать текстонезависимый алгоритм верификации диктора и улучшить его, используя текстозависимый подход
- Интегрирорвать средства для извлечения векторов признаков
- Интегрировать средства для выделения речевой активности
- Создать интерфейс для начального тестирования
- Протестировать систему верификации на речевом корпусе
- Определить оптимальные параметры разработанной системы

2 Процесс верификации

2.1 Цель

Допустим, что есть речевой сегмент Y и диктор S. Необходимо определить, был ли Y сказан диктором S или нет. Будем предполагать, что Y содержит речь только одного диктора.

Можем переформулировать задачу, определив гипотезы:

- H_0: Y сказан диктором S
- H_1: Y сказан не диктором S

Чтобы принять одну их представленных гипотез, мы должны рассмотреть отношение

$$\frac{p(Y|H_0)}{p(Y|H_1)} = \begin{cases} \geq \theta & \text{принимаем } H_0 \\ < \theta & \text{принимаем } H_1 \end{cases} \tag{1}$$

где $p(Y|H_i), i = 0, 1$, значение вероятностной функции плотности при условии гипотезы H_i, посчитанная для речевого сегмента Y, которая далее будет называться функцией правдоподобия, а θ пороговое значение, которое определяет выбор гипотезы. Значение θ рассчитывается и калибруется отдельно[5]. Система верификации диктора должна предоставить методику вычисления значения $p(Y|H_i)$, $i = 0, 1$, для любого речевого сегмента Y, для любого диктора S.

2.2 Схема процесса верификации

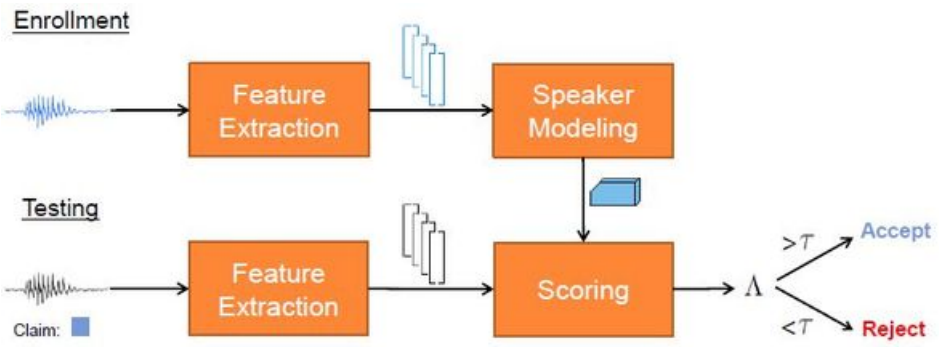

Рис. 1: Общая схема процесса верификации

Процесс верификации диктора можно разбить на две стадии:

- Построение эталонной модели
- Сравнение эталонной модели с тестовым произнесением

Этапы стадии построения эталонной модели:

1. Выделение речевой активности
2. Извлечение информативных признаков
3. Построение и адаптация модели диктора

Этапы стадии сравнения эталонной модели с тестовым произнесением:

1. Выделение речевой активности
2. Извлечение информативных признаков
3. Сравнение модели диктора и входящих данных
4. Принятие решения о верификации

3 Предобработка входных данных

3.1 Извлечение информативных признаков

Одним из основных компонентов системы по распознаванию или верификации диктора является процесс извлечения полезной информации из звукового сигнала. Применяя частотные фильтры к участку исходной записи получаем параметры, описывающие данный участок, из которых формируется вектор признаков[6]. Популярным решением для генерации векторов признаков является подход, использующий мел-кепстральные коэффициенты - MFCC (Mel-Frequency Cepstral Coefficients)[7]. Вектор признаков, характеризующий звуковой фрагмент, составляется из величины энергии, MFC коэффициентов, и их первой и второй производной. Рассматриваемые вектора признаков предоставляют достаточное количество информации о звуковом сигнале[8], поэтому было принято решение использовать их для последующего анализа и моделирования.

3.2 Выделение речевых данных

Для того, чтобы система по верификации диктора предоставляла стабильный результат для звуковых записей, сделанных в различных шумовых обстановках, необходимо провести этап предобработки входных данных. Применение детектора речевой активности предоставит возможность разметить данные на речевые и неречевые сегменты. Раздельная обработка описанных сегментов позволит улучшить результат верификации[9].

4 Построение модели диктора

4.1 Gaussian Mixture Model

Основываясь на опыте и результатах исследований, представленных в работах [4, 5, 10] для создания модели, характеризующей диктора и сказанную им фразу, был выбран метод, использующий вероятностную модель, построенную на гауссовских смесях (GMM, Gaussian Mixture Model). Преимущество данной модели заключается в способности точно описывать распределение векторов акустических признаков.

4.2 Описание модели GMM

GMM модель состоит из M компонент. Каждая компонента характеризуется: весом p_i, математическим ожиданием μ_i размерности $D \times 1$, матрицей ковариации Σ_i размерности $D \times D$, где $i = 1, \ldots, M$, а D - длина вектора признаков. Будем обозначать GMM модель говорящего λ, где

$$\lambda = \{p_i, \mu_i, \Sigma_i\} \quad i = 1, \ldots, M. \tag{2}$$

Функция плотности GMM вычисляется как взвешенная сумма плотностей M компонент её составляющих

$$p(\overrightarrow{x}) = \sum_{i=1}^{M} p_i b_i(\overrightarrow{x}) \tag{3}$$

где $\overrightarrow{x} \in X = \{\overrightarrow{x_1}, \ldots, \overrightarrow{x_T}\}$ - совокупность векторов признаков, построенных по произношениею Y; $i = 1, \ldots, M$; веса p_i удовлетворяют равенству $\sum\limits_{i=1}^{M} p_i = 1$.

Плотность каждой компоненты GMM вычисляем как функцию

$$b_i(\overrightarrow{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\overrightarrow{x} - \overrightarrow{\mu_i})^T \cdot \Sigma_i^{-1} \cdot (\overrightarrow{x} - \overrightarrow{\mu_i})\} \tag{4}$$

где μ_i математическое ожидание, а Σ_i матрица ковариации i-й компоненты GMM.

4.3 Настройка параметров GMM

Для настройки параметров μ_i и Σ_i, характеризующих конкретный речевой сегмент Y, воспользуемся итеративным алгоритмом EM (Expectation maximiztion)[10]. Необходимо максимизировать значение функции правдоподобия - $p(X|\lambda)$ для GMM модели $\lambda = \{p_i, \mu_i, \Sigma_i\}$, $i = 1, \ldots, M$ на тренировочных данных X. Например, для данных $X = \{\overrightarrow{x_1}, \ldots, \overrightarrow{x_T}\}$ функция правдоподобия вычисляется как:

$$p(X|\lambda) = \prod_{t=1}^{T} p(\overrightarrow{x_t}|\lambda). \tag{5}$$

где $p(\overrightarrow{x_t}|\lambda)$ вычисляется по формулам (3) и (4).

На практике используется не функция правдоподобия, а логарифм функии правдоподобия, усредненный на длину тестового произнесения:

$$\mathcal{L}(\lambda) = \frac{1}{T}\sum_{t=1}^{T}\log\left(\sum_{i=1}^{M} p_i b_i(\overrightarrow{x_t})\right) \tag{6}$$

где $b_i(\overrightarrow{x_t})$ вычисляется по формуле (4) с параметрами модели λ.

Начиная с базовой модели λ, переходим к модели $\overline{\lambda}$ так, что на каждом шаге $\mathcal{L}(\overline{\lambda}) \geq \mathcal{L}(\lambda)$. Для этого обновляем:

Веса:

$$\overline{p_i} = \frac{1}{T}\sum_{t=1}^{T} p(i|\overrightarrow{x_t}, \lambda) \quad i = 1, \ldots, M. \tag{7}$$

Математическое ожидание:

$$\overrightarrow{\overline{\mu_i}} = \frac{\sum\limits_{t=1}^{T} p(i|\overrightarrow{x_t}, \lambda)\overrightarrow{x_t}}{\sum\limits_{t=1}^{T} p(i|\overrightarrow{x_t}, \lambda)} \quad i = 1, \ldots, M. \tag{8}$$

Ковариационную матрицу:

$$\overline{\sigma_i}^2 = \frac{\sum\limits_{t=1}^{T} p(i|\overrightarrow{x_t}, \lambda)\overrightarrow{x_t}^2}{\sum\limits_{t=1}^{T} p(i|\overrightarrow{x_t}, \lambda)} - \overrightarrow{\overline{\mu_i}}^2 \quad i = 1, \ldots, M. \tag{9}$$

Где слагаемое из формул (7),(8),(9) вычисляется как:

$$p(i|\overrightarrow{x_t}, \lambda) = \frac{p_i b_i(\overrightarrow{x_t})}{\sum\limits_{k=1}^{M} p_k b_k(\overrightarrow{x_t})} \quad i = 1, \ldots, M. \tag{10}$$

5 Адаптация модели диктора

5.1 Universal Background Model

Допустим, что существует настроенная GMM модель диктора λ. Используя данную модель, мы сможем посчитать значение функции правдоподобия (5) при условии гипотезы H_0. В соотношении (1) фигурирует альтернативная гипотеза H_1. Для моделирования альтернативной гипотезы вводится UBM(Universal Background Model)[5].

UBM - это GMM модель, созданная на основе большого корпуса речи, отражающая информацию о речевых признаках группы людей. Используя UBM модель, вычисляется значение функции правдоподобия для H_1.

5.2 Адаптация модели диктора с использованием UBM

В этой главе описывается процесс, называемый MAP адаптацией модели диктора[5], использующий UBM. Проблема заключается в том, что полезной информации, извлеченной из короткой речевой фразы Y не хватит на создание полноценной модели. Поэтому в данном случае используется UBM модель, только не для моделирования альтернативной гипотезы, а для компенсации недостатка входных данных при создании модели диктора и сказанной им фразы.

Последовательность действий:

1. Создаем UBM на основе речевого корпуса, используя EM алгоритм
2. Извлекаем вектора признаков X по записанному речевому сегменту Y
3. Создаем адаптированную модель диктора

Процесс создания адаптированной модели диктора состоит из 2-х этапов:

1. Применяя формулы (7), (8), (9), вычисляем статистики, используя вектора признаков X и UBM модель λ. Полученные параметры обозначим как:
$$\hat{\lambda} = \{\hat{p}_i, \hat{\mu}_i, \hat{\Sigma}_i\}, \quad i = 1, \ldots, M$$
2. На основе параметров λ и новых параметров $\hat{\lambda}$ вычисляем результирующие параметры $\overline{\lambda}$.

Результирующие параметры $\overline{\lambda}$ находятся из следующих соотношений:

$$\overline{p_i} = \alpha_i \hat{p}_i + (1 - \alpha_i) p_i \tag{11}$$

$$\overline{\mu_i} = \alpha_i \hat{\mu}_i + (1 - \alpha_i) \mu_i \tag{12}$$

$$\overline{\sigma_i^2} = \alpha_i \hat{\sigma_i^2} + (1 - \alpha_i)(\sigma_i^2 + \mu_i^2) - \overline{\mu_i^2} \tag{13}$$
$$i = 1, \ldots, M$$

Параметр α_i, контролирующий баланс между λ и $\overline{\lambda}$ выбирается как:

$$\alpha_i = \frac{\hat{p}_i}{\hat{p}_i + r} \tag{14}$$

где значение $r = 16$ было подобрано экспериментально[5].

5.3 Hidden Markov Model

Для того, чтобы использовать информацию о временной структуре пароля диктора вводится Скрытая Марковская модель(HMM, Hidden Markov Model)[11], построенная с использованием адаптированной модели диктора. Таким образом модель будет совмещать в себе дикторскую и лингвистическую информацию.

5.4 Адаптация модели диктора с использованием HMM

Предположим, что имеется модель диктора λ, адаптированная при помощи UBM с использованием векторов признаков $X = \{\overrightarrow{x_1}, \ldots, \overrightarrow{x_T}\}$. Для извлечения дополнительной текстозависимой информации проводится двухэтапный процесс адаптации, использующий HMM.

На первом этапе адаптации вектора X разбиваются на N сегментов одинакового размера $\{seg_i\}, i = 1, \ldots, N$. Далее, для каждого сегмента seg_i независимо проводится настройка параметра веса модели λ по формулам (7), (11).

На втором этапе адаптации осуществляется пересегментация X на основе пути Витерби[11], вычисляемого для X, и матрицы переходов, рассчитываемой в зависимости от относительной длины сегментов[12]. На каждом новом сегементе seg_i проводится настройка параметра веса модели λ по формулам (7), (11). Второй этап повторяется до тех пор, пока процесс сегментации не стабилизируется.

6 Принятие решения о верификации

Для принятия решения о верификации диктора вычислим логарифм соотношения (1). Гипотезу H_1 характеризует UBM модель λ^{UBM} с параметрами $\{p_i^{UBM}, \mu_i^{UBM}, \Sigma_i^{UBM}\}$, $i = 1, \ldots, M$. Используя равенство (6) вычисляем $\mathcal{L}(\lambda^{UBM})$ на входных данных $X = \{\overrightarrow{x_1}, \ldots, \overrightarrow{x_T}\}$.

При помощи HMM адаптации извлекается информация о временной структуре фразы, представленная наборами весов p_j^k, где $j = 1, \ldots, M$, $k = 1, \ldots, \mathcal{N}$. Используя адаптированную модель λ^{GMM} с параметрами $\{p_i^{GMM}, \mu_i^{GMM}, \Sigma_i^{GMM}\}$, $i = 1, \ldots, M$ по матрице $X = \{\overrightarrow{x_1}, \ldots, \overrightarrow{x_T}\}$, разбитой на сегменты $\{seg_i\}, i = 1, \ldots, \mathcal{N}$ вычисляем логарифм значения функции правдоподобия для гипотезы H_0:

$$\mathcal{L}(\lambda^{HMM}) = \frac{1}{T} \sum_{t=1}^{T} \log \left(\sum_{i=1}^{M} p_i^k b_i(\overrightarrow{x_t}) \right) \tag{15}$$

где $b_i(\overrightarrow{x_t})$ вычисляется по формуле (4) с параметрами $\mu_i^{GMM}, \Sigma_i^{GMM}$, а p_i^k весовой параметр i-й компоненты GMM, настроенной на сегменте seg_k.

Результирующее значение $\mathcal{S}$ получается из формулы [12]:

$$\mathcal{S} = \mathcal{L}(\lambda^{HMM}) - \mathcal{L}(\lambda^{UBM}) \tag{16}$$

В соответствии с пунктом (1) значение $\mathcal{S}$ сравнивается с пороговым значением θ и принимается решение о верификации.

7 Реализация

Были выбраны вектора признаков размерности 42×1, содержащие значение энергии, 13 MFC коэффициентов, их первые и вторые производные. Для извлечения векторов использовалась библиотека VOICEBOX.[2]Представленная библиотека также использовалась для выделения речевой активности.

Для тестирования был выбран речевой корпус MIT[13]. В данном корпусе данные распределены на группы:

1. Зарегистрированные пользователи
2. Самозванцы

В записи данных участвовали мужчины и женщины, в регистрационной записи - 48 человек, в записи самозванцев - 40 человек. Регистрационная запись проводилась дважды. Каждый человек произнес 54 фразы из определенного списка. Для записи использовались 2 разных микрофона. Запись проводилась в 3-х различных шумовых обстановках.

Для создания UBM, состоящих из 128 и 64 компонент, использовались данные первой регистрационной записи.

Сравнение модели диктора со входными данными считалось целевым, если модель и сказанная фраза принадлежали одному диктору и сказанная фраза совпадала с требуемым паролем. Иначе сравнение считалось нецелевым. Ожидается, что система верификации должна подтверждать целевое сравнение и отклонять нецелевое.

Было сформировано 2 протокола тестирования. Протокол A содержал целевые сравнения и нецелевые сравнения различных дикторов. Протокол B содержал целевые сравнения, нецелевые сравнения различных дикторов, нецелевые сравнения одинаковых дикторов с различающимися паролями. По каждому из протоколов проводилось 20000 сравнений.

По оси X (Рис.2) в процентном отношении представлена вероятность ошибки первого рода, вероятность ошибочно отвергунть гипотезу H_0 из (1). По оси Y в процентном отношении представлена вероятность ошибки второго рода, вероятность ошибочно принять гипотезу H_0. Значение, при котором совпадают вероятности ошибки первого и второго рода, называется равновероятной ошибкой (EER, Equal Error Rate). EER описывает точность системы верификации диктора.

Для UBM состоящей из 128 компонент, для протокола A, на 6819 целевых и 13181 нецелевых сравнениях EER $= 15.15\%$ (Рис.2).

Для UBM состоящей из 128 компонент, для протокола B, на 5141 целевых и 14859 нецелевых сравнениях EER $= 17.12\%$ (Рис.3).

Для UBM состоящей из 64 компонент, для протокола A, на 6819 целевых и 13181 нецелевых сравнениях EER $= 18.21\%$.

Для UBM состоящей из 64 компонент, для протокола B, на 5141 целевых и 14859 нецелевых сравнениях EER $= 19.04\%$.

[2]VOICEBOX - открытая библиотека по обработке речи, реализованная на MATLAB. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (дата обращения: 11.04.2014)

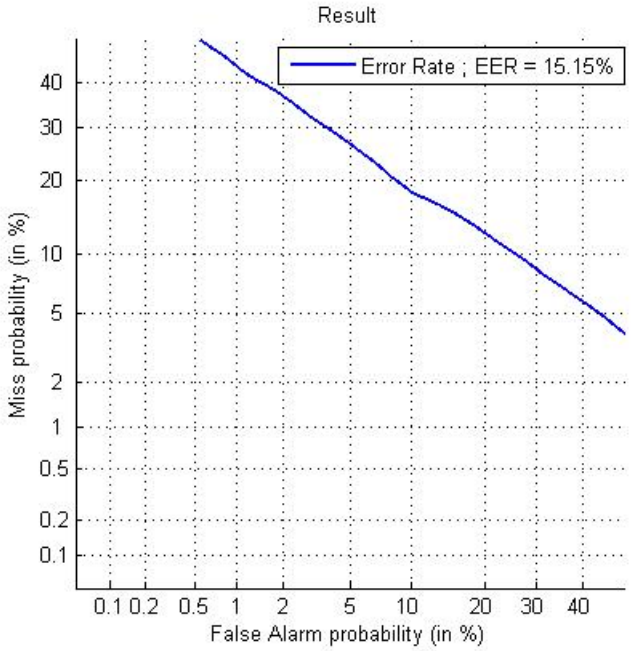

Рис. 2: Результаты тестирования системы верификации по протоколу A для UBM из 128 компонент

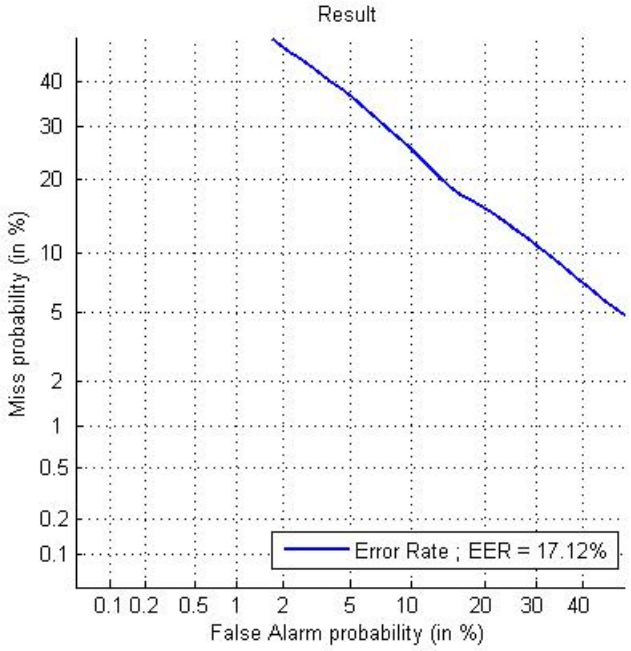

Рис. 3: Результаты тестирования системы верификации по протоколу B для UBM из 128 компонент

8 Результаты

В рамках курсовой работы были получены следующие результаты:

- Изучены основные алгоритмы верификации диктора
- Реализован текстозависимый алгоритм верификации диктора
- Интегрированы средства для извлечения векторов признаков
- Интегрированы средства для выделения речевой активности
- Создан интерфейс для тестирования
- Проведено тестирование на речевом корпусе MIT[13]

Список литературы

[1] Fauve B., Evans N., Pearson N., Bonastre J.-F., Mason J.S.: Influence of task duration in text-independent speaker verification// Annual Conference of the International Speech Communication Association. 2007. P.794–797.

[2] Bonastre, J.-F, Wils, F., Meignier, S.: ALIZE, a free Toolkit for Speaker Recognition// IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 1. 2005. P.737 - 740.

[3] Sadjadi, S. O., Slaney, M., Heck L.: MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker-Recognition Research, November 2013. http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2013-11/IdentityToolbox/

[4] Larcher, A.B., Mason J.F., John S.D.: Constrained Temporal Structure For Text-dependent Speaker Verification// Digital Signal Processing. Vol. 23. 2013. P.1910-1917.

[5] Reynolds, D. A., Quatieri, T. F., Dunn, R. B.: Speaker Verification using adapted Gaussian mixture models// Digital Signal Processing. Vol. 10. 2000. P.19–41.

[6] Young S., Evermann G., Gales M., Hain T., Kershaw D.: The HTK Book 3. 1999. P.77-78, 191-193, 217-222.

[7] Benesty, J., Sondhi, M.M., Huang, Y.: Handbook of Speech Processing. Springer. 2008.

[8] O'Shaughnessy, D.: Formant Estimation and Tracking// Springer Handbook of Speech Processing. 2008. P.213–227.

[9] Droppo, J., Acero, A.: Environmental Robustness// Springer Handbook of Speech Processing. 2008. P.662–663.

[10] Reynolds, D. A., Rose, R. C.: Robust text-independent speaker identification using Gaussian mixture speaker models// IEEE Transactions on Audio, Speech and Language Processing. Vol.3. 1995. P.72–83.

[11] Huang, X., Acero, A., Hon, H-W, Spoken Language Processing: A Guide to Theory, Algorithm and System Development. 2001. 377-413.

[12] Larcher, A.B., Mason J.F., John S.D.: From GMM to HMM for Embedded Password-Based Speaker Recognition// European Signal and Image Processing Conference. 2008

[13] Woo R.H., Park A., Hazen T.: The MIT Mobile Device Speaker Verification Corpus: Data collection and preliminary experiments// Proceedings of IEEE Odyssey. 2006