Система идентификации дикторов по голосу для конкурса NIST SRE 2010

д.т.н. Матвеев Ю.Н., Симончик К.К.
ООО «Центр Речевых Технологий», Санкт-Петербург, Россия
matveev@speechpro.com, simonchik@speechpro.com

The speaker identification system for the NIST SRE 2010
Matveev Yu.N., Simonchik K.K.

ABSTRACT (ENG)

The speaker identification system, submitted for the NIST Year 2010 Speaker Recognition Evaluation (SRE), is described.

Keywords: *Biometry, speaker identification, voice recognition, speech features, pitch, formants, GMM, SVM, NIST.*

Аннотация

Приведено описание системы идентификации дикторов по голосу, разработанной для конкурса по оцениванию систем распознавания дикторов NIST SRE 2010.

Ключевые слова: *Биометрия, идентификация диктора, распознавание по голосу, основной тон, форманты, СГР, SVM, NIST.*

1. ВВЕДЕНИЕ

Системы идентификации (распознавания, верификации) дикторов по голосу относятся к классу биометрических систем, достоинством которых является то, что они чаще всего не требуют дополнительного оборудования для регистрации голоса и могут быть реализованы с использованием телефонных сетей или устройств ввода-вывода разных типов (микрофонов).

Область использования такого рода приложений обширна:

– автоматическая идентификация подозреваемого в телефонном канале;

– автоматическая верификация клиентов при удаленном доступе по телефонному каналу;

– обработка речевых баз данных;

– криминалистические исследования.

В данной работе представляется описание текстонезависимой системы автоматической идентификации дикторов по голосу, разработанной ООО «Центр Речевых Технологий» для участия в международном конкурсе по оцениванию систем распознавания дикторов NIST SRE 2010.

В профессиональной среде NIST SRE (Speaker Recognition Evaluation) называют неофициальным чемпионатом мира по голосовой идентификации. Начиная с 1996 года, этот конкурс ежегодно проводится американским Национальным Институтом Стандартов и Технологий (англ. National Institute of Standards and Technology, NIST). Его основная цель – оценить уровень существующих технологий и определить перспективные направления развития индустрии. Регулярно в конкурсе принимают участие ведущие компании, университеты и лаборатории со всего мира. В 2010 году в конкурсе участвовало 46 научных команд.

Первой особенностью оценивания NIST SRE 2010 являлось использование баз речевых данных, собранных по различным каналам связи и в акустике помещений, и характеризующимися широким диапазоном отношений сигнал-шум и уровней реверберации (см. рисунок 1). Точка на скаттерограмме обозначает присутствие в речевом корпусе фонограммы с определенным уровнем шума (SNR, dB) и уровнем реверберации (Reverberation time).

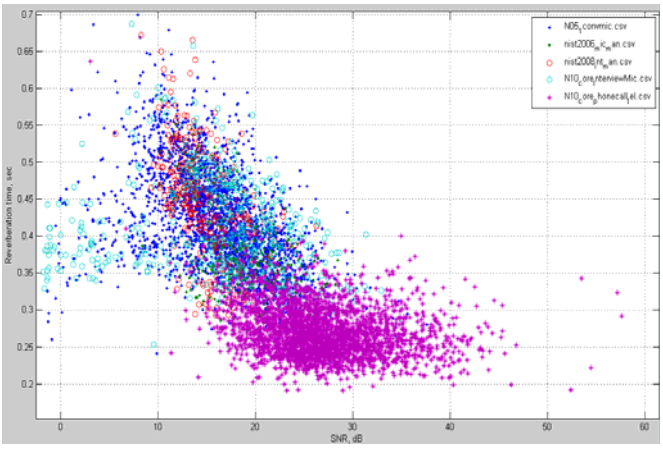

Рисунок 1: Скаттерограмма корпусов речевых данных NIST различных годов: фиолетовый цвет – сотовый корпус 2010 г.; остальные – корпуса речевых данных в акустике помещений 2005-2010 г.г.

Второй особенностью NIST SRE 2010, в сравнении с предыдущими конкурсами, являлось то, что в этом году организаторы задали новую функцию минимизации ошибки идентификации, суть которой состояла в крайне высокой стоимости ошибки ложного пропуска нецелевого диктора:

$$DCF = FR + 999\,FA,$$

где *FR* - false rejection error rate (вероятность ошибки ложного отклонения);

FA - false acceptance error rate (вероятность ошибки ложного пропуска);

Введение новых значений весов параметров требует значительных объемов данных для калибровки порога принятия решения системы идентификации, в связи с тем,

что количество попыток идентификации нецелевого диктора должно быть достаточно большим для статистически устойчивой оценки *DCF*.

Точность калибровки особенно важна при применении голосовой идентификации в реальных условиях и зачастую играет критическую роль, так как позволяет максимально точно адаптироваться под прикладные задачи.

В данной работе описывается система, которая показала один из лучших результатов по качеству идентификации, в том числе, заняла первое место по уровню калибровки среди коммерческих систем.

2. МЕТОДЫ ИДЕНТИФИКАЦИИ ДИКТОРА

Принцип работы системы идентификации диктора основан на выделении из фонограмм речи и последующем попарном сравнении биометрических признаков (содержащихся в голосе индивидуальных, идентификационно значимых, признаков личности).

Выделение и сравнение биометрических признаков производится с использованием различных языко- и текстонезависимых методов идентификации дикторов по голосу. Система распознавания диктора называется текстонезависимой, если она не содержит информации о том, что именно диктор будет произносить (система обучается и тестируется на произвольных речевых данных).

На данный момент наиболее распространённым подходом к решению задач текстонезависимой идентификации является подход на основе использования моделей гауссовых смесей (англ. Gaussian Mixture Models, GMM) [1]. В качестве речевых признаков в подавляющем числе систем идентификации используются мэл-частотные кепстральные коэффициенты (англ. Mel-Frequency Cepstral Coefficients, MFCC) [2]. Модель голоса диктора представляет собой аппроксимацию распределения используемых речевых признаков смесью гауссовых распределений (GMM-модель).

Значения равновероятной ошибки (англ. Equal Error Rate, EER) принятия чужого и отбрасывания своего диктора для метода на основе MFCC-GMM зависят от длительности сравниваемых речевых фрагментов и могут достигать величины ~4–5%.

Однако, при относительно высокой точности идентификации, по сравнению с другими популярными методами, такими как спектрально-формантный (СФ) метод [3] и метод идентификации на основе статистик основного тона (СОТ) [4], метод MFCC-GMM предъявляет высокие требования к качеству сигнала, обладает сильной зависимостью от вида обучающего материала, а также требует относительно больших временных затрат на выделение биометрических признаков. Сравнительные характеристики перечисленных методов приведены в таблице 1 (количество знаков «+» отражает степень зависимости метода от параметров сигнала).

Таблица 1

Метод	Параметры сигнала		
	Продолжи-тельность	Качество сигнала	Физическое и эмоциональное состояние
СФ	+++	++	+
СОТ	++	+	++++
MFCC-GMM	+++	++++	++

Для улучшения показателей системы идентификации на основе GMM-моделей голосов дикторов обычно используются:

– канало-компенсация [5],

– классификатор на основе машины опорных векторов (англ. support vector machine, SVM),

Следует отметить, что совместное использование GMM-SVM моделей является на сегодняшний день доминантным в задачах верификации/идентификации/распознавания дикторов по голосу [1].

3. ОПИСАНИЕ СИСТЕМЫ

3.1 Структура системы

В системе автоматической идентификации личности по голосовым признакам в естественной речи осуществляется сравнение одной или нескольких записей (моделей) голоса неизвестного диктора с одной или несколькими записями (моделями) голоса известного диктора. В результате такого сравнения определяется, насколько похож голос неизвестного диктора на голос известного и, следовательно, принадлежат ли записи речи одному человеку или разным людям.

Если тестируемая фонограмма речи диктора может не принадлежать ни одному из кандидатов, то в систему дополнительно вводится фоновая модель (модель «самозванца»), и схема такой системы, называемой системой идентификации дикторов на открытом множестве, имеет вид, представленный на рисунке 2.

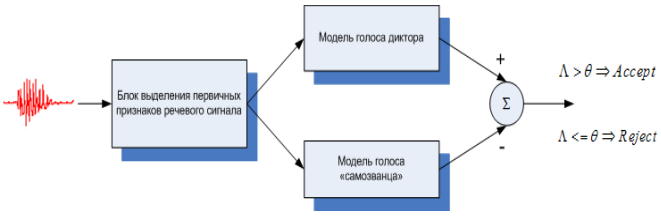

Рисунок 2: Структурная схема системы идентификации дикторов по голосу на открытом множестве.

Для каждого диктора-кандидата проводится сравнение речевого сигнала с моделью голоса данного диктора. Однако, кроме этого, вычисляется также вероятность того, что речевой сигнал принадлежит какому-нибудь другому диктору – на основании так называемой универсальной фоновой модели (англ. Universal Background Model, UBM) или модели «самозванца» (англ. impostor) на рисунке 2, которая описывает некоторые усредненные характеристики всех дикторов по используемой речевой базе.

3.2 Описание системы

Представленная на конкурс NIST SRE 2010 система состояла из 6 различных гендеро- и канало-зависимых подсистем. Подсистемы адаптировались для различных каналов получения фонограмм:

- телефонного;

- микрофонного;

- смешанного (телефон-микрофон).

Кроме того, в рамках одного канала производилось дополнительное деление подсистем на две гендеро-зависимые (для женских и мужских голосов) подсистемы.

В качестве обучающих данных были взяты речевые базы NIST SRE прошлых лет (2004, 2006 и 2008 гг.) общим объемом более 20 тыс. фонограмм.

Для повышения надежности системы в качестве дополнительных речевых признаков были использованы линейно-частотные кепстральные коэффициенты (англ. Linear-Frequency Cepstral Coefficients, LFCC) [2], что обеспечило повышение качества идентификации в микрофонном канале.

Результирующее решение ("decision") по обоим признакам (MFCC и LFCC) получалось путем вычисления «обобщенного решения» ("fusion"), получаемого методом взвешенного голосования, когда результату работы каждой подсистемы присваивается некоторый вес (см. рисунок 3). Для определения этих весов на этапе обучения системы использовался инструментарий собственной разработки. Обучение и точная калибровка системы производилась на речевой базе NIST SRE 2005.

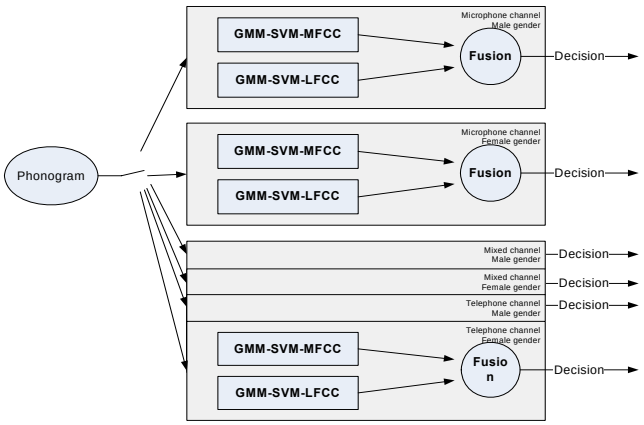

Рисунок 3: Структура системы идентификации диктора ООО «Центр Речевых Технологий».

GMM-модели голосов дикторов обучались методом максимального правдоподобия Фогта [5] с использованием компенсации канальных искажений. Размерность пространства собственных каналов для разных подсистем варьировалось от 50 до 80. В качестве классификатора использовался SVM с классической zt-нормализацией [5], которая представляла собой нормирование выходной дистанции SVM по случайным произнесениям дикторов из речевой базы объемом 1000-2000 фонограмм.

4. СМЕШИВАНИЕ ПОДСИСТЕМ И СИСТЕМ

4.1 Основные характеристики систем и подсистем

На конкурс NIST SRE 2010 были предоставлены три системы (SVID-1, SVID-2, SVID-3), которые отличаются корпусами речевых данных, используемых для обучения и настройки систем.

4.1.1 Primary system (SVID-1)

Базовая (primary) система является комбинацией двух подсистем, каждая из которых строилась на отдельных наборах речевых признаков:

– Первая подсистема: на базе 39-мерных векторов признаков, составленных из 13 MFCC-коэффициентов, дополненных их первыми и вторыми производными. Для каждого из векторов применялась процедура вычитания кепстрального среднего (CMS).
– Вторая подсистема: на базе 39-мерных векторов признаков, составленных из 13 LFCC-коэффициентов, дополненных их первыми и вторыми производными. Для каждого из векторов применялась процедура вычитания кепстрального среднего (CMS).

Каждая из подсистем, в свою очередь, имеет 6 гендеро- и канало-зависимых UBM. При обучении UBM использовались 1024-компонентные гауссовы смеси. База обучения UBM состояла из речевых корпусов Switchboard II Phases 2&3, Switchboard Cellular Parts 1&2, NIST SRE 2004, 2006&2008, из которых отбирались фонограммы дикторов, имеющих по 5–10 сессий записи их речи. Характеристики базы обучения представлены в таблице 2.

Таблица 2

Пол	Общее кол-во	Каналы		
		тел-тел	мик-мик	тел-мик
муж	дикторов	788	158	280 (153 тел+ 158 мик)
	фонограмм	6546	1516	2290
жен	дикторов	1042	203	371 (201 тел+ 203 мик)
	фонограмм	8589	1955	3050

Импостеры отбирались из РБД NIST SRE 2006, 2008. Характеристики базы импостеров представлены в таблице 3.

Таблица 3

Пол	Общее кол-во	Каналы		
		тел	мик	тел-мик
муж	дикторов	1070	1450	1000
жен	дикторов	1227	2236	1000

Для сегментации дикторов использовалась информация из ASR (Automatic Speech Recognition) транскрипции, предоставленной NIST.

Для получения обобщенного по всем подсистемам результирующего решения использовался собственной разработки инструментарий для смешивания по критерию минимизации функции стоимости DCF.

4.1.2 Secondary system (SVID-2)

Вторичная (secondary) система отличается использованием на этапе обучения UBM речевой базы NIST SRE 2005 вместо NIST SRE 2008.

4.1.3 Secondary system (SVID-3)

Еще одна вторичная (secondary) система была сформирована путем комбинирования первичной (SVID 1) и вторичной систем (SVID 2).

4.2 Смешивание подсистем

Обобщенное решение по всем подсистемам было основано на методе взвешенного голосования:

$$d(x) = \sum_{i=1}^{S} \alpha_i d_i(x) + \Theta$$

где $d_i(x)$ - выходное значение i-й подсистемы, α_i - весовой коэффициент для i-й подсистемы, Θ - пороговое значение, S - число подсистем.

Калибровка общего решения производилась на речевой базе NIST SRE 2005, которая не использовалась для обучения базовых GMM-SVM подсистем.

В связи с тем, что оценка *DCF* при высокой стоимости ошибки ложного пропуска не является статистически устойчивой, оптимизация коэффициентов α_i производилась простейшим методом перебора. При этом в качестве функции минимизации ошибки использовалась непосредственно функция *DCF*.

5. РЕЗУЛЬТАТЫ

В таблицах 4-5 приведены характеристики систем идентификации дикторов ООО «Центр Речевых Технологий» до конкурса и представленной на конкурс NIST SRE 2010.

Как следует из приведенных данных, было обеспечено:

– повышение точности идентификации (понижение EER) в 3-4 раза;

– робастность идентификации в различных условиях (шумы, реверберация);

– сохранение достигнутых параметров точности и робастности при кросс-канальных сравнениях.

Таблица 4

Каналы	Система до конкурса EER, %	Система после конкурса EER, %
Микрофон-микрофон (различные микрофоны)	15-18	6,0
Микрофон-телефон (различные телефонные каналы)		4,9
Телефон-телефон (различные телефонные трубки и каналы)		5,0

Таблица 5 - Смешанный корпус микрофон – GSM-канал

Система до конкурса	Система после конкурса

Длит-ность, сек	16	32	80	Длит-ность, сек	17	29	77
16	15,2	14,0	14,0	16	8,0	6,3	3,8
32		12,9	11,4	32		4,5	2,7
80			8,9	80			1,3

Исходя из официально предоставленных NIST материалов (см. рисунок 4), система идентификации дикторов ООО «Центр Речевых Технологий» заняла на конкурсе NIST SRE 2010:

- 2-е место по уровню калибровки (1-е место среди коммерческих компаний);
- 7-е место по фактической стоимости (actual cost) технологии - официальной метрике NIST (2-е место среди коммерческих компаний).

На рисунке 4 используются следующие обозначения:

Min DCF – минимальное значение DCF, порог определяется NIST;

Act DCF – реальное значение DCF, порог определен участником;

Min DCF / Act DCF – степень калибровки системы

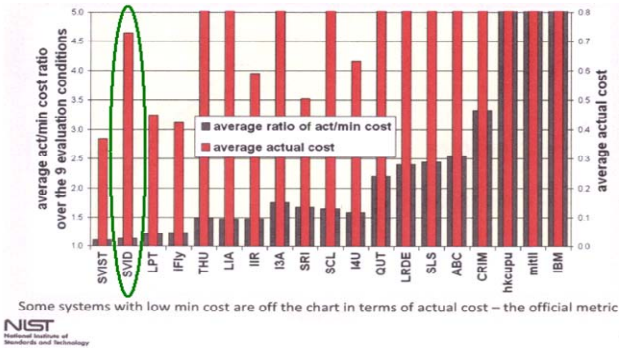

Рисунок 4: Фактическая стоимость (actual cost) детектирования системы по всем 9-ти DET-графикам.

6. ЗАКЛЮЧЕНИЕ

В рамках подготовки к конкурсу NIST SRE 2010 была произведена модернизация системы идентификации дикторов ООО «Центр Речевых Технологий», что обеспечило значительное повышение точности идентификации (понижение EER) в 3-4 раза, повышение робастности идентификации в различных условиях (шумы, реверберация) и повышение скорости системы при построении голосовых моделей по фонограмме.

В настоящее время предложенный в рамках NIST SRE 2010 подход к идентификации дикторов используется в системе VoiceNet ID ведения и автоматизации национального фоноучета Мексики [7].

7. СПИСОК ЛИТЕРАТУРЫ:

[1] Bimbot F. et al. A Tutorial on Text-Independent Speaker Verification. - EURASIP Journal on Applied Signal Processing, 2004, №4, p.p. 430–451.

[2] Reynolds D. Experimental evaluation of features for robust speaker identification. – IEEE Trans. On Speech and Audio Processing, 1994, vol. 2, №4, p.p. 639-643.

[3] Коваль С.Л., Лабутин П.В., Раев А.Н. Патент РФ 2230375 от 10.06.2004 «Метод распознавания диктора и устройство для его осуществления».

[4] Коваль С.Л., Лабутин П.В., Малая Е.В., Прощина Е.А. Идентификация дикторов на основе сравнения статистик основного тона голоса // В сб. трудов XV международной научной конференции «Информатизация и информационная безопасность правоохранительных органов». – М.: Академия управления МВД России, 2006. С. 324-327.

[5] Vogt R., Baker B., Sridharan S. Modeling session variability in text-independent speaker verification // In Proc. Eurospeech, Lisbon, Portugal, Sept. 2005, pp. 3109–3112.

[6] Rosenberg A. et al. Cepstral channel normalization techniques for HMM-based speaker verification // In Proc. ICSLP-94, pp. 1835-1838.

[7] Тимофеев А.В. Распределённая система фоноучёта «VoiceNet ID». – Речевые технологии, 2009, №2, с. 69-73.

Об авторах

Юрий Николаевич Матвеев – д.т.н., руководитель отдела верификации и идентификации диктора ООО «Центр Речевых Технологий», Санкт-Петербург.

Электронный адрес matveev@speechpro.com.

Константин Константинович Симончик – научный сотрудник отдела верификации и идентификации диктора ООО «Центр Речевых Технологий», Санкт-Петербург.

Электронный адрес simonchik@speechpro.com.