МРНТИ 16.31.21, 28.23.15
УДК 57.087.1: 004 (574)

СИСТЕМАТИЧЕСКИЙ ОБЗОР И АНАЛИЗ ОСОБЕННОСТЕЙ ИДЕНТИФИКАЦИИ ПО ГОЛОСУ

О.Ж. МАМЫРБАЕВ[1], А.С. КЫДЫРБЕКОВА[1,2], А.Т. АХМЕДИЯРОВА[1], М. ТҰРДАЛЫҰЛЫ[1], Н.О. МЕКЕБАЕВ[1,2]

[1]Институт информационных и вычислительных технологий
[2]Казахский Национальный университет им. аль-Фараби

Аннотация: *Идентификация по голосу – это процесс идентификации говорящего по данному высказыванию путем сравнения голосовой биометрии высказывания с теми моделями высказывания, которые были сохранены заранее. Технологии идентификации по голосу получили новое направление благодаря достижениям в области искусственного интеллекта и широко используются в различных областях. Извлечение признаков является одним из наиболее важных аспектов идентификации по голосу, который существенно влияет на процесс и производительность идентификации. Этот систематический обзор проводится для выявления, сравнения и анализа различных подходов, методов и алгоритмов извлечения признаков для идентификации по голосу, чтобы предоставить справочную информацию о подходах извлечения признаков для приложений идентификации по голосу и будущих исследований. В ходе исследования были рассмотрены модели: основанные на шаблонах, основанные на векторном квантовании, динамическом переносе времени, модель гистограмм, стохастические модели, модели гауссовой смеси и скрытая Марковская модель, основанные на Mel-частотных кепстральных коэффициентах, генеративное или векторное квантование, дискриминационные модели (обычно с использованием методов машинного обучения, таких как SVM и ANN). Это исследование показало, что текущая тенденция исследования идентификации заключается в разработке надежной универсальной структуры идентификации по голосу для решения важных проблем идентификации по голосу, таких как адаптивность, сложность, многоязычное распознавание и устойчивость к шуму. Результаты, представленные в этом исследовании, основаны на прошлых публикациях, цитатах и количестве реализаций, причем цитаты являются наиболее актуальными. Эта статья также представляет общий процесс идентификации по голосу.*

Ключевые слова: *биометрия, биометрическая аутентификация, голосовая аутентификация, речевые технологии, голосовая биометрия, идентификация по голосу, распознавание по голосу*

SYSTEMATIC REVIEW AND ANALYSIS OF THE PECULIARITIES OF IDENTIFICATION BY VOICE

Abstract*: Voice identification is the process of identifying a speaker by a given utterance by comparing voice biometrics of a utterance with those utterance models that were saved in advance. Voice identification technologies have gained a new direction due to advances in artificial intelligence and are widely used in various fields. Character extraction is one of the most important aspects of voice identification, which significantly affects the identification process and performance. This systematic review is conducted to identify, compare and analyze various approaches, methods and algorithms for extracting features for voice identification, to provide background information on character retrieval approaches for voice identification applications and future research. The study examined models: based on patterns, based on vector quantization, dynamic time transfer, histogram model, stochastic models, Gaussian mixture models and hidden Markov model, based on Mel-frequency cepstral coefficients, generative or vector quantization, discriminatory models (usually using machine learning methods such as SVM and ANN).This study showed that the current trend of identification research is to develop a robust, universal voice identification structure for solving important voice identification*

problems, such as adaptability, complexity, multilingual recognition, and resistance to noise. The results presented in this study are based on past publications, quotations and the number of implementations, the quotes being the most relevant. This article also presents the general process of voice identification.

Keywords: *biometrics, biometric authentication, voice authentication, speech technology, voice biometrics, Identification by voice, voice recognition*

ДАУЫС АРҚЫЛЫ ИДЕНТИФИКАЦИЯЛАУДЫҢ ЕРЕКШЕЛІКТЕРІНЕ ТАЛДАУ ЖӘНЕ ЖҮЙЕЛІК ШОЛУ

Аңдатпа: *Дауыс бойынша идентификациялау деп алдын ала сақталған дыбысталу моделі мен сөйлеушінің дыбысталған деректерінің дауыстық биометриясын салыстыру арқылы сөйлеушіні идентификациялау процесін айтамыз. Дауыс бойынша идентификациялау технологиялары жасанды интеллекті саласының жетістіктері нәтижесінде жаңа бағыт алды және әртүрлі салаларда кеңінен қолданысқа ие болып келеді. Белгілерді бөліп алу дауыс бойынша идентификациялаудың ең маңызды аспектілерінің бірі болып табылады және идентификациялаудың өнімділігі мен орындалу процесіне айтарлықтай әсер етеді. Аталған жүйелік шолу дауыс бойынша идентификациялауға арналған белгілерді бөліп алудың әртүрлі тәсілдерін, әдістері мен алгоритмдерін айқындау, салыстыру және талдау үшін және дауыс бойынша идентификациялаудың қосымшалары мен болашақ зерттеулерге арналған белгілерді бөліп алу тәсілдері туралы ақпараттық анықтама ұсыну үшін жүргізілген. Зерттеу барысында, үлгілерге, векторлық кванттауға және уақытты динамикалық ауыстыруға негізделген модельдер, гистограмм, стохастикалық, гаустық үлестіру және Мел-жиіліктегі кепстральді коэффициенттерге, генеративті немесе векторлық кванттауға және дискриминативтік модельдерге (әдетте SVM және ANN секілді машиналық оқыту әдістерін қолдану арқылы) негізделген жасырын Марков модельдері қарастырылған. Бұл зерттеу, идентификациялау зерттеулерінің қазіргі беталысы, дауыс бойынша идентификациялаудың бейімділік, күрделілік, көптілді тану және шуылға төзімділік секілді маңызды мәселелерін шешуге арналған идентификациялаудың сенімді әмбебап құрылымын құру екенін көрсетті. Осы жұмыста берілген зерттеу нәтижелері өткен басылымдарға, цитаталарына және оларды жүзеге асырулардың санына негізделеді, оның ішінде цитаталар ең өзектісі болып табылады. Мақалада дауысты идентификациялаудың жалпы процесі де ұсынылады.*

Түйінді сөздер: *биометрия, биометриялық аутентификация, дауыстық аутентификация, сөйлеу технологиялары, дауыстық биометрия, дауыс бойынша идентификациялау, дауыс бойынша тану*

Введение

Биометрия остается незаменимой там, где необходимо обеспечить безопасность доступа к физическим объектам и информационным ресурсам. Биометрические технологии успешно применяются в правоохранительной деятельности, гражданской регистрации, в области безопасности банковских обращений, инвестирования, в вопросах охраны здоровья и во многих сферах деятельности. Наиболее часто используемыми биометрическими характеристиками являются отпечатки пальцев, форма лица, радужная оболочка глаза, голос, подпись, геометрия руки. Мы не можем сказать, что та или иная характеристика является лучше остальных. Чтобы выбрать подходящий биометрический метод идентификации нужно учитывать такие факторы, как область его применения, требуемый уровень безопасности, целевую установку (верификация или идентификация), ожидаемое число пользователей, практичность и другие. Биометрия – это разработка статистических и математических методов, применяемых к анализу данных проблемы в биологических науках. Внедрение этой технологии обеспечивает новую безопасность подходов к компьютерным системам. Идентификация и проверка – это два способа использования биометрических данных для идентификации личности. Биометрия относится к использованию физических или физиологических, биологических или поведенческих характе-

ристик для установления личности человека. Эти характеристики уникальны для каждого человека и остаются частично неизменными в течение жизни [1].

Хотя биометрия не является идеальным решением, она предлагает несколько преимуществ по сравнению с подходами, основанными на знаниях и владении, в том смысле, что нет необходимости запоминать что-либо, биометрический факт в силу того, что эти атрибуты очень трудно подделать и требуют, чтобы атрибуты не могли быть потеряны, переданы или украдены, предлагают лучшую защиту – наличие подлинного пользователя при предоставлении доступа к определенным ресурсам. Вдохновленные развитием системы Бертиллона в 1883 году и системы элементарного распознавания отпечатков пальцев сэра Джона Гальтона в 1903 году, исследовательское сообщество посвятило свои усилия открытию нескольких биометрических методов. Любой физиологический или поведенческий атрибут может быть квалифицирован как биометрический признак, если только он не удовлетворяет таким критериям, как:

– универсальность: присуща всем людям;

– различие: дискриминация среди населения;

– инвариантность: выбранный биометрический признак должен демонстрировать неизменность во времени;

– собираемость: легко собирать с точки зрения сбора, оцифровки и выделения признаков из популяции;

– производительность: относится к наличию ресурсов и наложению реальных ограничений в плане сбора данных и гарантии достижения высокой точности;

– приемлемости: готовности населения передать этот атрибут в систему распознавания и обходу:

– склонность к подражанию или подражанию в случае мошеннических атак на систему распознавания.

Классификация распознавания по голосу. Речь – это универсальная форма общения. Распознавание по голосу (РГ) – это процесс идентификации говорящего по голосовым особенностям данной речи. Это отличается от распознавания речи, когда процесс идентификации ограничен содержанием, а не говорящим. Процесс РГ основан на выявлении и извлечении уникальных характеристик речи говорящего. Характеристики голоса человека также называют голосовой биометрией. Система РГ используется для идентификации и различения говорящих и извлечения уникальных характеристик, которые могут использоваться для проверки или аутентификации пользователя. Идентификация по голосу (ИГ) известна как процесс идентификации говорящего по данному высказыванию путем сравнения голосовой биометрии данного образца говорящего. Когда голос используется для авторизации, он называется «Проверка речи». Ключевой областью применения РГ является безопасность и криминалистика. Системы РГ также используются в качестве замены для пароля и других процессов аутентификации пользователя (голосовой пароль). Криминалистика применяет ИГ для сравнения голоса человека, полученного, например, по телефону или другими зарегистрированными доказательствами. Этот процесс также называется обнаружением динамика. Наиболее важным аспектом использования систем ИГ является автоматизация таких процессов как перенаправление писем клиентов в нужный почтовый ящик, распознавание собеседников в обсуждении, предупреждение рамок подтверждения дискурса при смене говорящего, проверка того, зачислен ли клиент в каркас и так далее. Эти системы ИГ могут работать без знания образца голоса клиента, так как они полагаются только на идентификацию входящего оратора из существующей базы данных ораторов.

Наш систематический обзор ограничен ИГ как одним из основных типов систем ИГ. Извлечение признаков является одним из важных аспектов ИГ, который существенно влияет на качество ИГ. В частности, выбор подходящих методов извлечения признаков играет жизненно важную роль, поскольку идентификация осуществляется путем срав-

нения уникальных характерных признаков голосового ввода. Поэтому целью этой статьи является систематический обзор литературы по различным подходам к извлечению признаков ИГ с целью:

1) определить подходы к извлечению существенных признаков за последние шесть лет;

2) представить систематический обзор исследований подходов извлечения признаков для ИГ;

3) классифицировать различные подходы к извлечению признаков и дать рекомендации, основанные на исследовании.

Приложения идентификация по голосу могут быть классифицированы на два типа и показаны на рисунке 1:

– первый тип зависит от наличия голосовых отпечатков в базе данных, которая далее классифицируется на две категории, а именно закрытый набор и открытый набор. В закрытом режиме вход тестового динамика сравнивается с речевыми отпечатками существующих ораторов в базе данных, и обнаруживается ближайшее совпадение [2]. Следовательно, закрытый ИГ гарантирует результат, хотя он может не быть точным говорящим. С другой стороны, в открытой системе ИГ голосовой отпечаток входного динамика сравнивается с базой данных для «точного соответствия»; ввод отклоняется, если совпадение не найдено (Reynolds, 2002);

– второй тип приложений ИГ основан на уровне пользовательского контроля, который также известен как процесс проверки динамиков. Эта категория ИГ также подразделяется на две категории: текстозависимое и текстонезависимое. В текстозависимой говорящий должен произносить те же фразы или слова, которые ранее использовались для обучения [3,4], в то время как в последней категории входной голосовой контент для печати может отсутствовать в обучении set [5, 6, 7].

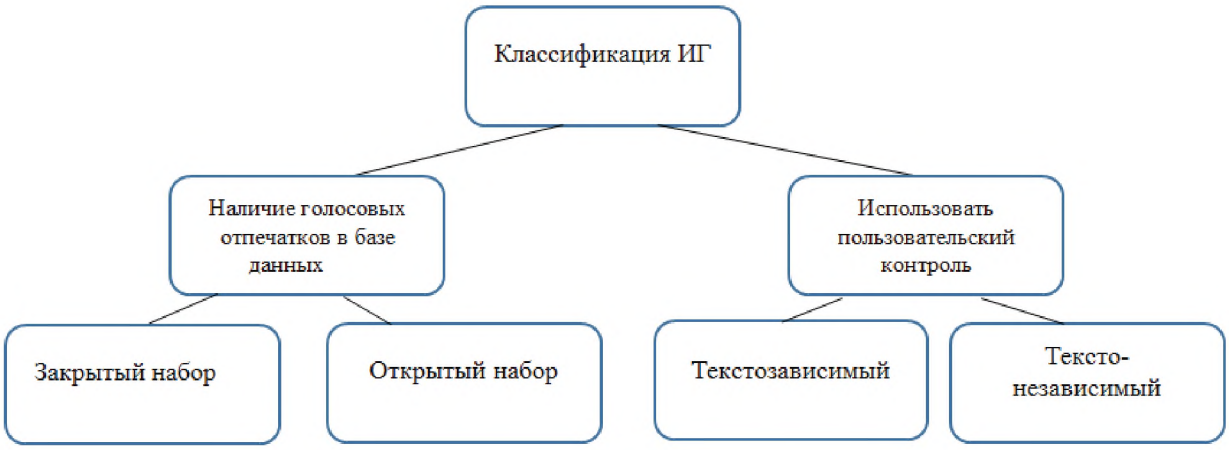

Рис. 1 – Классификация приложений идентификации по голосу

Систематический обзор проводится с использованием методологии систематического обзора литературы, предложенный Китченом и Чартерс (2007), который подробно описан в разделе методологии. В этом обзоре мы представили различные подходы извлечения признаков, которые использовались в процессах идентификации докладчиков, и представили систематический обзор исследований этих подходов. Следует отметить ключевые компоненты систем ИГ (которые подробно описаны в следующих разделах), такие как параметризация (т.е. извлечение признаков), моделирование динамиков, сопоставление с образцом и метод оценки, которые также являются основными компонентами для РГ. Этот систематический обзор объяснил все компоненты ИГ, но больше внимания было уделено извлечению признаков.

Поскольку РГ можно рассматривать как проблему распознавания образов, для систем РГ используются различные подходы искусственного интеллекта. Недавние реализации глубокого обучения для РГ подчеркивают сложность, связанную с РГ, которая требу-

ет особого внимания по сравнению с традиционными проблемами распознавания образов в целом и проблемами распознавания речи в частности [8].

Существующие обзорные работы и опросы по распознаванию ораторов можно в целом разделить на три категории.

Первая категория – это всесторонние обзоры по ИГ, в которых рассматривается литература по общим процессам ИГ и различным категориям ИГ.

Вторая категория в основном сосредоточена на типах статистического и машинного обучения, используемых в качестве классификаторов ИГ, например, Farrell, Mammone и Assaleh (1994); Ларчер, Ли, Ма и Ли (2014); Липпманн (1989) [9,10,11]. Эта категория обзоров ИГ в основном попадает под классификацию и исследования машинного обучения, где имеется значительное количество доступной литературы. С точки зрения идентификации говорящего, единственная работа, которая конкретно обсуждала ИГ и ее процессы – это краткий обзор, представленный Сидоровым, Шмиттом, Заблоцким и Минкером (2013), в котором было объяснено и сравнено несколько общих методов ИГ [12].

Третья категория опросов по распознаванию ораторов касается подходов к извлечению признаков в ИГ. Одним из последних исследований по извлечению характеристик ИГ является Disken, Tüfekçi, Saribulut и Çevik (2017), в которых основное внимание уделяется методам выделения надежных специфических характеристик динамиков на основе профилей шума, эмоций и несоответствия каналов [13]. Другим примером является Rao and Sarkar (2014), в котором представлено упрощенное объяснение систем верификации динамиков на основе моделей и функций [14]. Был также представлен краткий обзор Chavan and Chougule (2012), в котором кратко определены и объяснены особенности распознавания по голосу [15]. Еще одна недавняя работа, в которой подчеркивается важность использования подходов глубокого обучения для выделения функций при идентификации говорящих – это Тирумала и Шахамири (2017) [16].

Другими примерами являются обзор по анализу ИГ, моделированию и извлечению характеристик, представленный Jayanna и Prasanna (2009), и опрос по оценке акустических характеристик ИГ, основанный на экспериментальных оценках (Lawson et al., 2011) [17,18].

Ни одна из существующих работ по обзору ИГ не проводилась с использованием методологии систематического обзора. Кроме того, эти работы не были ограничены определенным периодом и вместо этого были подготовлены общие отчеты о ходе исследований в области ИГ. Наконец, они не представили подробный анализ процесса идентификации говорящего и подчеркнули важность выделения признаков в РГ. Таким образом, в литературе есть пробел в предоставлении систематической ссылки на подходы извлечения признаков РГ. В этой статье делается попытка восполнить этот пробел путем систематического обзора, представляющего подробный обзор большинства статистических подходов и подходов к извлечению новых возможностей машинного обучения. Мы также подробно представили процесс идентификации докладчика. В частности, в данном документе собрана соответствующая информация и обсуждены технологии РГ, о которых с особым вниманием к методам выделения признаков РГ сообщалось в литературе с 2013 по 2018 год. Уместно отметить, что методы оценки не входят в сферу данного исследования. Основной вклад этой статьи заключается в объединении всех связанных реализаций РГ в одном месте, что послужит справочным материалом для исследователей РГ. Кроме того, этот документ может быть использован для предложения критериев выбора конкретной модели извлечения признаков для реализации систем РГ.

Процесс идентификации по голосу. В этом разделе представлены классификации распознавания по голосу, а затем описывается процесс идентификации по голосу.

С точки зрения исследования, ИГ может быть разделен на категории в соответствии с выполненным действием или на основе облас-

ти исследования, как показано на рисунке 2. В частности, области ИГ, основанные на действии:

1) идентификация по голосу (ИГ) для идентификации неизвестного пользователя на основе его голосовых отпечатков (Daoudi, Jourani, Andre, & Aboutajdine, 2011; Wu & Lin, 2011). Это процесс сравнения одного голосового профиля пользователя со многими профилями и поиска наилучшего или точного соответствия [19,20];

2) проверка по голосу (или Аутентификация) – это процесс проверки личности пользователя с помощью его голосовых отпечатков, когда говорящий утверждает, что он является конкретным пользователем (Jiang, Gao & & Han, 2009). Это матч один на один [21];

3) диаризация по голосу – это идентификация голоса человека из данной группы, когда говорящий говорит (Anguera et al., 2012; Poignant, Besacier, & Quénot, 2015) или же является сочетанием методов сегментации и кластеризации дикторов [22,23]. РГ отличается от диаризации динамика: в системе РГ вход обычно представляет собой однопользовательский голос, и цель состоит в том, чтобы сопоставить речевые характеристики с профилем динамика из источника данных. Тем не менее, при динаризации говорящего система произносит смесь высказываний различных пользователей, в то время как цель системы состоит в том, чтобы идентифицировать речь конкретного пользователя и определить, когда он говорит;

4) распознавание спикера используется для сохранения анонимности пользователей. Обычно используется для скрытия личности пользователей, когда их личность должна быть скрыта при сохранении акустической информации от речи (Jin, Toth, Schultz & Black, 2009; Justin et al., 2015; Pobar & Ipsic, 2014) [24,25,26];

5) обнаружение голосовой активности – это процесс определения существования человеческой речи (Haigh & Mason, 1993; Ram, Segura, Ben, De La Torre, & Rubio, 2004) [27,28].

Основанные на исследованиях области идентификации по голосу являются следующими:

1) моделирование речевых технологий – это процесс идентификации и привязки уникального идентификатора к голосовым отпечаткам отдельных говорящих, чтобы отличать их от других выступающих, представленных в базе данных [29];

Рис. 2 – Основные направления исследований для распознавания голоса

2) параметризация речи – это процесс вычисления набора параметров из небольшой части голосовых отпечатков динамика, который описывает свойства динамика или речевого сигнала [30];

3) методы сопоставления и оценки шаблонов используются для сравнения шаблонов, представленных в голосовых отпечатках входного динамика, с шаблонами, извлеченными из различных динамиков для соответствия уникальным характерным признакам. Затем каждому совпадению сходства присваивается оценка, которая определяет точность идентификации говорящего. Процесс и фазы РГ описаны ниже.

В целом, система идентификации по голосу проходит две основные фазы: фазу обучения, которая также называется регистрацией, и фазу соответствия, где регистрация проверяется на соответствие.

Фаза регистрации начинается с приема входных сигналов моделирования речи и предварительной обработки и нормализации данных. Следующим шагом является извлечение признаков, обеспечивающее параметры речевого сигнала таким образом, который понятен системе. Извлеченные функции также могут нуждаться в нормализации до начала процесса обучения. Процесс обучения может включать как автономное обучение (алгоритм обучения, фоновое моделирование, адаптация модели), так и онлайн-обучение (адаптация модели). Результаты этапа обучения – это модели динамиков, которые сохраняются для использования на следующем этапе. В разделе распознования по голосу: регистрация голосов объясняет этот этап более подробно. Задача фазы согласования состоит в том, чтобы сопоставить речевой сигнал, полученный от говорящего, который должен быть идентифицирован (то есть тестовый громкоговоритель), с моделями громкоговорителей, сохраненными во время фазы регистрации, чтобы идентифицировать говорящего, произносящего речь. Как и в первом этапе, входной сигнал должен быть предварительно обработан, нормализован, а его характеристики должны быть извлечены. Далее, характеристики тестового динамика сравниваются с моделями обученных динамиков, которые ищут соответствие. Затем следует вычисление показателя сходства и его нормализация. В разделе распознавания по голосу предоставляет дополнительную информацию о фазе сопоставления.

Распознавание по голосу: регистрация голосов. Эта фаза инициируется параметризацией речи, при которой речевой ввод предварительно обрабатывается и нормализуется перед извлечением признаков. Извлеченные речевые функции также могут нуждаться в нормализации перед созданием и сохранением речевых отпечатков или моделей для динамика. Процесс параметризации речи.

1. Ввод речевого сигнала: необходимо учитывать следующие параметры входных сигналов:

– источник речи: определить, является ли речевой сигнал живым предметом или это записанная речь;

– язык: разные языки могут выделять разные типы речевых особенностей, которые влияют на производительность РГ;

Большинство систем РГ в литературе разработано для английского языка, в литературе также представлены системы РГ не на английском языке. Например: японский (Kawakami, Wang, Kai, & Nakagawa, 2014), тайский (Tanprasert & Achariyakulporn, 2000), испанский (Luengo et al., 2008) [31,32,33,34]. Кроме того, есть некоторые известные индийские языки, такие как ассамский (Sarma & Sarma, 2013a), маратхи (Jawarkar, Holambe & & Basu, 2012) и хинди (Jawarkar, Holambe & & Basu, 2015) [35, 36,37].

Существует также несколько подходов к идентификации многоязычных говорящих, таких как [38]:

– устройство захвата речи: различные типы устройств захвата речи записывают речь по-разному, поскольку они имеют разные типы датчиков, уровень возможностей захвата и чувствительность.

Например, некоторые типы микрофонов предназначены для захвата речевых сигналов для конкретной среды. Существуют так-

же микрофоны, которые предназначены для определенной цели, такие как портативные микрофоны, микрофоны с шумоподавлением, компьютерные микрофоны и микрофоны, встроенные в телефоны или смартфоны;

– шум окружающей среды: различные профили шума, такие как фоновый шум, шум окружающей среды и эхо в помещении могут нарушать входной сигнал и существенно влиять на производительность систем обработки речи. На чувствительность микрофона также влияет шум; например, в тихих условиях высокочувствительный микрофон может захватывать не только оригинальный звук динамика, но и сигналы реверберации [39]. Методы помехоустойчивости или умные помещения могут использоваться для снижения воздействия шума [40];

– изменчивость речи: манера говорящего, например, скорость, громкость, болезнь, возраст, эмоции, время суток (например, утренний или вечерний голос) и т.д. Также могут изменяться речевые особенности. Примером является исследование, проведенное Сахидуллой, Чакроборти и Саха (2011) для изучения влияния шести эмоциональных состояний (нейтральное, счастливое, грустное, злое, отвращение и страх) в системах РГ. Определение настроения говорящего также рассматривалось в работах Ахмеда, Кенкеремата и Станковича (2015). [41,42];

2) предварительная обработка речи: этот процесс имеет дело с любыми помехами или сбоями, которые могут повлиять на извлечение признаков. Он в основном пытается удалить шумы и пробелы в тишине. Это важно, потому что шумы и промежутки молчания во входных данных обладают весьма нестационарными характеристиками, которые могут вызвать ложную идентификацию [43];

3) нормализация: это помогает устранить любые изменения, такие как изменчивость и изменчивость между сессиями, которые могут вызывать колебания речевых характеристик за счет потери некоторых характеристик. Эта изменчивость между сессиями происходит из-за изменений в среде записи, условиях передачи, фоновом шуме и изменениях в голосах ораторов. Одинаковые высказывания говорящего не могут повторяться подобным образом для каждого испытания, то есть высказывание, записанное в одном сеансе, сильно коррелирует с записями в другом сеансе.

Обычной практикой нормализации является использование банков фильтров. Существует два типа методов нормализации, а именно на основе параметров и расстояния (или сходства). Первый тип нормализации доказал свою эффективность для уменьшения влияния длинных спектральных вариаций и линейных каналов. Текстовые системы распознавания говорящих, которые имеют достаточно длинные высказывания, применяют процесс, называемый спектральным выравниванием (так называемое слепое выравнивание), который представляет собой процесс устранения помех или шума с помощью нелинейного сигнала отклика фразы. В этом процессе общее среднее значение кепстральных коэффициентов суммарных высказываний вычитается из усредненных значений кепстральных коэффициентов отдельного кадра;

4) извлечение характеристик: это процесс представления акустического сигнала в качестве специфических акустических характеристик. Функции выбираются так, чтобы наилучшим образом представить акустическую характеристику сигналов для различных типов систем обработки речи. Это подробно обсуждается в следующем разделе.

Моделирование громкоговорителей – это следующий этап фазы обучения РГ, на которой модели громкоговорителей обучаются с использованием извлеченных акустических характеристик громкоговорителей. Эффективность моделей динамиков отражается в точной идентификации динамиков с целью минимизации частоты ошибок. Существует три типа методов моделирования: классические подходы, параметрические подходы, основанные на парадигме обучения, и гибридные методы, в которых применяются методы машинного обучения.

Иерархия для различных подходов к моделированию динамиков:

1. Классический подход. 2. Парадигма обучения. 3. Гибридный подход.

1. Классический подход имеет два типа моделей; основанные на шаблонах модели, основанные на векторном квантовании, динамическом переносе времени или моделях гистограмм, или стохастические модели, основанные на модели гауссовой смеси (GMM) или скрытой марковской модели (HMM).

2. Модели обучающей парадигмы (параметрические подходы) могут быть генеративными (например, GMM или векторное квантование) или дискриминационными моделями (обычно с использованием методов машинного обучения, таких как SVM и ANN).

3. Гибридные подходы представляют собой комбинацию вышеуказанных моделей, таких как GMM-HMM, ANN-HMM и т.д. Кроме того, некоторые подходы, такие как векторное квантование и GMM, могут быть классифицированы как в классической, так и в обучающей парадигмах.

Параметрические подходы соответствуют некоторому распределению обучающих данных путем поиска параметров распределения, которые максимизируют требуемый критерий. Непараметрические подходы, с другой стороны, делают минимальные предположения о распределении признаков.

Процесс обучения может быть в автономном режиме или онлайн. Оффлайн модели обучения требуют фиксированного хранилища спикеров. Для обучения дискриминационных моделей необходимо использовать данные докладчиков из хранилища в качестве отрицательных выборок для распознавания. Кроме того, подходы автономного обучения используются в универсальном фоновом моделировании для уникальной адаптации ораторов на основе векторов их характеристик. Таким образом, для обучения в автономном режиме требуются все известные модели громкоговорителей.

Универсальное фоновое моделирование используется в онлайн-обучении. Тем не менее, модели динамиков принимаются «в режиме онлайн» из обучающих данных докладчиков в режиме реального времени. В частности, модели применяются в режиме реального времени, что делает его более надежным для идентификации неизвестных ораторов (то есть моделей, которые не представлены в обучении). Эта адаптация также помогает ускорить вторую фазу РГ, что объясняется в следующем разделе.

Распознавание по голосу: этап идентификации или соответствия. Этап идентификации начинается с параметризации речи, которая была подробно описана в предыдущем разделе. Затем следует этап сопоставления идентификаторов, на котором функции, извлеченные из неизвестных высказываний говорящего, передаются системе для идентификации говорящего. Оценки сходства (то есть вероятность) также получают, сравнивая данное входное высказывание с любой моделью динамика, сохраненной в системе. Сопоставление с образцом является вероятностным в стохастических моделях, где результаты рассчитываются в форме условных вероятностей; с другой стороны, шаблонные модели используют детерминированные подходы.

Соответствующая часть этой фазы обычно применяет алгоритмы сопоставления с образцом, которые указаны выше в одной из моделей.

Он отвечает за идентификацию говорящего путем сопоставления обученных моделей говорящего с использованием извлеченных функций из неизвестного высказывания говорящего. Чтобы определить наилучшее соответствие, процесс идентификации сравнивает высказывание с несколькими моделями громкоговорителей или голосовыми отпечатками.

На протяжении всего планирования определены цели предлагаемой систематическому обзору литературы и выполнены необходимые оценки:

1) определение потребности в систематическом обзоре литературы: в процессе планирования определено, что в области распознавания по голосу не было представлено ни одного недавнего систематического обзора литературы, в котором основное внимание уделяется моделированию динамиков и параметризации речи. Аналогичным об-

разом, в 2010 году был опубликован обзорный документ, в котором рассматривается переход от векторных моделей распознавания по голосу к супервекторным парадигмам между 1980 и 2013 годами. Тем не менее этот обзор не проводился систематически. Более того, недавно было введено несколько новых методов в контексте извлечения признаков распознавания по голосу, которые необходимо систематически пересматривать. В этом документе основное внимание уделяется заполнению этого пробела путем определения и обобщения эволюции выделения признаков в РГ, особенно с 2011 года. Этот систематический обзор также предоставляет исчерпывающую справочную информацию для исследователей РГ;

2) формулировка вопросов: сформулированы следующие вопросы для этого обзора:

– каковы критерии для оптимальных функций и как функция параметров соответствия определяется для процесса извлечения признаков в РГ?

– какие подходы и алгоритмы извлечения признаков используются в процессе идентификации голоса?

– какой самый популярный и успешный подход к извлечению признаков за последние шесть лет?

3) выбор соответствующих ресурсов: процесс расследования проводился в соответствии с руководящими принципами систематического обзора различных методов; процесс поиска проводился с фиксированной датой начала и окончания с указанием месяца и года, как это рекомендовано в Stapic, Lo, Cabot, de Marcos Ortega и Strahonja (2012). Во время поиска были рассмотрены популярные цифровые ресурсы, такие как цифровая библиотека IEEE Xplore, ScienceDirect, цифровая библиотека ACM, Google Scholar, DBLP и Springer Verlag.

Процесс проведения обзора объясняется следующим:

1) идентификация исследования: при первоначальном поиске с использованием распознавания по голосу в качестве ключевого слова без каких-либо фильтров было получено от ScienceDirect, Scopus из цифровой библиотеки. Ресурсы с объединенными результатами из разных источников, таких как Google Scholar и записей соответственно.

В результате изменения ключевого слова на идентификацию голоса и применения фильтра ограничения времени на 2013-2018 гг. было выявлено 78 соответствующих статей. Этот этап дополнительно поясняется на этапе «выбора»;

2) стратегия сбора первичных исследований: значимость отборочных статей, определенных на последнем этапе, была изучена путем тщательного изучения разделов реферата, методологии, обсуждения и результатов каждого документа. Для классификации применялись следующие критерии, и документы, которые не соответствовали им, были исключены:

– в документе обсуждалась идентификация докладчиков для академических или научных электронных библиотек;

– в документе обсуждалась идентификация по голосу с целью предоставления рекомендаций для книг или статей;

– в документе обсуждалась идентификация по голосу с целью предоставления рекомендаций по методам распознавания по голосу для академических или научных электронных библиотек или рекомендаций по научным документам;

– набор данных эксперимента относится к процессу распознавания голоса, который включает в себя извлечение признаков как его часть;

– метод идентификации по голосу, обсуждаемый в статье, создан для академической и научной аудитории;

– метод идентификации по голосу, обсуждаемый в документе, был создан для любого практического использования в биометрических приложениях.

3) извлечение данных и обобщение: для каждого документа следовали правилам систематического обзора различных методов для извлечения и синтеза данных. Работы были включены в короткий список на основе ответов на вопросы исследования, предоставленные исследовательскими материалами.

Для данной ситуации эти признанные темы и критерии представили классификации, приведенные в сегменте выводов и результатов.

Процессор извлечения данных может быть определен как объем классификации, предоставленной на протяжении извлечения данных, а также объем, предоставленный на этапе синтеза данных. Систематический обзор различных методов с рекомендациями не дает подробного и четкого описания процесса извлечения данных. Следовательно, выбрано тривиальное извлечение данных, что привело к записи цитат, которые были лишь незначительно переписаны; на этапе синтеза такие цитаты были ранней классификации. В этом разделе показаны частоты количества раз, когда каждый субъект распознается в различных источниках, в которых каждому событию соответствовал одинаковый вес. Такие частоты отражают только то, как часто данный вопрос выделяется в различных статьях. Тем не менее, невозможно определить, насколько это важно.

Заключение

В данной статье представлен систематический обзор различных методов извлечения признаков и алгоритмов идентификации голосов. Были использованы научные рекомендации для извлеченных публикаций из различных источников. Представлен общий процесс РГ с последующим подробным обзором различных процессов извлечения признаков. Также обсуждалась важность выявления значимых особенностей и их влияние на точность идентификации говорящего.

В этом исследовании было рассмотрено около 557 публикаций с 2011 года с использованием методологии систематического обзора. Первоначально составили базу данных, извлекая 557 статей, за которыми следовал четырехуровневый процесс фильтрации и применяя критерии, такие как период, язык и т.д. В результате тщательного изучения этих статей было получено 65 публикаций, которые были изучены для данного обзора литературы. Рассматривая эти статьи, можно увидеть, что большинство подходов к извлечению признаков применяют форму MFCC независимо от алгоритма, используемого для фазы классификации голосов.

В литературе не рекомендуется применять комплексный подход для построения строгих рекомендаций. Следовательно, здесь использовались рекомендации, основанные на подходах, включенных в различные научные статьи. Этот обзор литературы может служить ресурсом для извлечения признаков в отношении идентификации по голосу.

Исходя из выводов этого исследования, рекомендуются следующие направления будущих исследований.

1. Следует отметить, что не существует обобщенного подхода к обобщению универсальных признаков, и данное исследование подчеркивает его необходимость. Однако многие исследовательские проблемы направлены на то, чтобы сохранить компромисс между точностью идентификации по голосу и устойчивостью к шуму, который все еще требует большого внимания.

Рекомендуется в будущих исследованиях изучить вопрос о разработке надежной универсальной структуры для идентификации по голосу, которая решает важные проблемы распознавания по голосу.

2. В будущих исследованиях следует дополнительно изучить их применение в распознавании по голосу как в качестве экстрактора признаков, так и в качестве классификатора, а также выяснить, как они могут внести вклад в универсальную структуру идентификации по голосу.

3. Была проведена ограниченная работа по методам идентификации неконтролируемых динамиков, которые включают создание моделей динамиков на основе функций, извлеченных из немаркированных данных.

4. Аналогичным образом, извлечение отличительных признаков динамика из неполных, подделанных или поврежденных данных не было должным образом изучено, несмотря на их широкое применение в криминалистике и восстановлении данных. Необходимо систематически изучать, как работают существующие подходы распознавания по голосу, когда им дают подделанные данные, и исследовать методы для улучшения их производительности.

ЛИТЕРАТУРА

1. A. Jain L.Hang and S. Pankanti. “Can multi-biometrics improve performance,” Proceedings of Auto ID, 59-64, 1999.
2. Dutta, M., Patgiri, C., Sarma, M., & Sarma, K. K. (2015). Closed-set text-independent speaker identification system using multiple ANN classifiers. In Proceedings of the 3rd international conference on frontiers of intelligent computing: Theory and applications (FICTA) 2014 (pp. 377–385).
3. Islam, M. R., & Rahman, M. F. (2009). Improvement of text dependent speaker identification system using neuro-genetic hybrid algorithm in office environmental conditions. International Journal of Computer Science Issues, 1, 42–48.
4. Kekre, H. B., Athawale, A., & Desai, M. (2011). Speaker identification using row mean vector of spectrogram. In Proceedings of the international conference and workshop on emerging trends in technology (pp. 171–174).
5. Boujelbene, S. Z., Mezghanni, D. B. A., & Ellouze, N. (2009). Robust text independent speaker identification using hybrid GMM-SVM System. International Journal of Digital Content Technology and its Applications, 3, 103–110.
6. Revathi, A., & Venkataramani, Y. (2009). Text independent composite speaker identification/ verification using multiple features. In 2009 WRI World congress on computer science and information engineering: 7 (pp. 257–261).
7. Verma, G. K. (2011). Multi-feature fusion for closed set text independent speaker identification. In International conference on information intelligence, systems, technology and management (pp. 170–179).
8. Richardson, F., Reynolds, D., & Dehak, N. (2015a). Deep neural network approaches to speaker and language recognition. IEEE Signal Processing Letters, 22, 1671–1675.
9. Farrell, K. R., Mammone, R. J., & Assaleh, K. T. (1994). Speaker recognition using neural networks and conventional classifiers. IEEE Transactions on speech and audio processing, 2, 194–205.
10. Hong Kong, China: IEEE. Larcher, A., Lee, K. A., Ma, B., & Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and RSR2015. Speech communication, 60, 56–77.
11. Lippmann, R. P. (1989). Review of neural networks for speech recognition. Neural computation, 1, 1–38.
12. Sidorov, M., Schmitt, A., Zablotskiy, S., & Minker, W. (2013). Survey of automated speaker identification methods. In 2013 9th international conference on intelligent environments (IE) (pp. 236–239).
13. Dis¸ ken, G., Tüfekçi, Z., Saribulut, L., & Çevik, U. (2017). A review on feature extraction for speaker recognition under degraded conditions. IETE Technical Review, 34, 321–332.
14. Rao, K. S., & Sarkar, S. (2014a). Robust speaker verification: A review. In Robust speaker recognition in noisy environments (pp. 13–27).
15. Chavan, M., & Chougule, S. (2012). Speaker features and recognition techniques: A review. International Journal of Computational Engineering Research, 2, 720–728.
16. S.S. Tirumala et al. / Expert Systems With Applications 90 (2017) 250–271.
17. Nagaraja, B. G., & Jayanna, H. S. (2012). Multilingual speaker identification with the constraint of limited data using multitaper MFCC. In International conference on security in computer networks and distributed systems (pp. 127–134).
18. Lawson, A., Vabishchevich, P., Huggins, M., Ardis, P., Battles, B., & Stauffer, A. (2011). Survey and evaluation of acoustic features for speaker recognition. In Acoustics, speech and signal processing (ICASSP), 2011 ieee international conference on (pp. 5444–5447).

19. Daoudi, K., Jourani, R., Andre, O. R. e. g., & Aboutajdine, D. (2011). In Speaker identification using discriminative learning of large margin GMM: 6 (pp. 300–307).
20. Shih, P.-Y., Lin, P.-C., Wang, J.-F., & Lin, Y.-N. (2011). Robust several-speaker speech recognition with highly dependable online speaker adaptation and identification. Journal of network and computer applications, 34, 1459–1467.
21. Jiang, S., Frigui, H., & Calhoun, A. W. (2015). Speaker identification in medical simulation data using fisher vector representation. In 2015 IEEE 14th international conference on machine learning and applications (ICMLA) (pp. 197–201).
22. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing, 20, 356–370 6.
23. Poignant, J., Besacier, L., & Quénot, G. (2015). Unsupervised speaker identification in TV broadcast based on written names. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23, 57–68.
24. Jin, Q., Toth, A. R., Schultz, T., & Black, A. W. (2009). Speaker de-identification via voice transformation (pp. 529–533).
25. Justin, T., Struc, V., Dobrisek, S., Vesnicer, B., Ipsic, I., & Mihelic, F. (2015). In Speaker de-identification using diphone recognition and speech synthesis: 4 (pp. 1–7).
26. Pobar, M., & Ipsic, I. (2014). Online speaker de-identification using voice transformation. In 2014 37th International convention on information and communication technology, electronics and microelectronics (mipro) (pp. 1264–1267).
27. Haigh, J. A., & Mason, J. S. (1993). Robust voice activity detection using cepstral features. In 1993 IEEE Region 10 conference on proceedings. computer, communication, control and power engineering (TENCON'93): 3 (pp. 321–324).
28. Ram, i,r. J., Segura, J. e. C., Ben, i,t. C., De La Torre, A., & Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. Speech Communication, 42, 271–287.
29. Beigi, H. (2011). Speaker Modeling. In Fundamentals of speaker recognition (pp. 525–541).
30. Ganchev, T. (2011). Contemporary methods for speech parameterization. pp. 233–236.
31. Kawakami, Y., Wang, L., Kai, A., & Nakagawa, S. (2014). Speaker identification by combining various vocal tract and vocal source features. In International conference on text, speech, and dialogue (pp. 382–389).
32. Kawakami, Y., Wang, L., & Nakagawa, S. (2013). Speaker identification using pseudo pitch synchronized phase information in noisy environments. In 2013 Asia-Pacific on signal and information processing association annual summit and conference (APSIPA) (pp. 1–4).
33. Tanprasert, C., & Achariyakulporn, V. (2000). Comparative study of GMM, DTW, and ANN on Thai speaker identification system. Sixth international conference on spoken language processing, ICSLP 2000 / INTERSPEECH 2000.
34. Luengo, I., Navas, E., Sainz, I. n. a., Saratxaga, I., Sanchez, J., & Odriozola, I. (2008). Text independent speaker identification in multilingual environments. In Proceedings of the international conference on language resources and evaluation, LREC 2008.
35. Sarma, M., & Sarma, K. K. (2013a). Speaker identification model for Assamese language using a neural framework. In The 2013 international joint conference on neural networks (IJCNN) (pp. 1–7).
36. Jawarkar, N. P., Holambe, R. S., & Basu, T. K. (2012). Text-independent speaker identification in emotional environments: A classifier fusion approach. In Frontiers in Computer Education (pp. 569–576).
37. Jawarkar, N. P., Holambe, R. S., & Basu, T. K. (2015). Effect of nonlinear compression function on the performance of the speaker identification system under noisy conditions. In Proceedings of the 2nd International Conference on Perception and Machine Intelligence (pp. 137–144).

38. Nagaraja, B. G., & Jayanna, H. S. (2012). Multilingual speaker identification with the constraint of limited data using multitaper MFCC. In International conference on security in computer networks and distributed systems (pp. 127–134).
39. Wang, L., Zhang, Z., & Kai, A. (2013). Hands-free speaker identification based on spectral subtraction using a multi-channel least mean square approach. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 7224–7228
40. Busso, C., Hernanz, S., Chu, C.-W., Kwon, S.-i., Lee, S., & Georgiou, P. G. (2005). Smart room: Participant and speaker localization and identification. IEEE International Conference on Acoustics, Speech, and Signal Processing: 2. IEEE. Campbell, J. P. (1997). Speaker recognition: A tutorial. Proceedings of the IEEE, 85, 1437–1462 6.
41. Sahidullah, M., Chakroborty, S., & Saha, G. (2011). Improving performance of speaker identification system using complementary information fusion. In Proceedings of 17th international conference on advanced computing and communications (pp. 182–187).
42. Ahmed, M. Y., Kenkeremath, S., & Stankovic, J. (2015). Socialsense: A collaborative mobile platform for speaker and mood identification. In Wireless sensor networks: 8965 (pp. 68–83).
43. Farhood, Z., & Abdulghafour, M. (2010). Investigation on model selection criteria for speaker identification. In 2010 International symposium in information technology (ITSim): 2–6 (pp. 537–541).