Информационные процессы, Том 12, №1, стр. 1-30

====== ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ТЕХНИЧЕСКИХ ======
====== И СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ СИСТЕМАХ =======

РАСПОЗНАВАНИЕ ЛИЧНОСТИ ПО ГОЛОСУ: АНАЛИТИЧЕСКИЙ ОБЗОР

В.Н.Сорокин, В.В.Вьюгин, А.А.Тананыкин

Институт проблем передачи информации, Российская академия наук, Москва, Россия
Поступила в редколлегию 25.01.2012

Аннотация. Задача распознавания диктора по его голосу была поставлена более 40 лет тому назад, и исследования в этой области все еще продолжаются. Решение этой задачи может найти применение в криминалистике, радио-разведке, контр-разведке, антитеррористическом мониторинге, обеспечение безопасности доступа к физическим объектам, информационным и финансовым ресурсам. В зависимости от конкретной задачи различают верификацию и идентификацию диктора. В первом случае пользователь указывает свой идентификатор, и требуется либо подтвердить его или отказать в подтверждении. Во втором случае необходимо идентифицировать диктора среди множества других дикторов.

В большинстве работ для распознавания диктора используются параметры в виде коэффициентов кепстра, который вычисляется по огибающей спектра, полученного через преобразование Фурье, с помощью гребенки фильтров, либо по передаточной функции речевого тракта, найденной методом линейного предсказания. В дополнение к коэффициентам кепстра используются также их первые и вторые разности по времени. Преимущество такого подхода заключается в вычислительной простоте, а также в том, что в кепстре отражаются индивидуальные характеристики голосового источника и анатомия речевого тракта. Вместе с тем, различительная способность такого описания ограничена, и поэтому значительные усилия сконцентрированы на разработке решающих правил. Наиболее популярны методы гауссовых смесей (GMM) и опорных векторов (SVM). Используются также искусственные нейронные сети и скрытые Марковские модели (HMM).

С целью сравнения различных методов распознавания диктора введен показатель равной ошибки (EER), определяющий ошибку распознавания при условии равенства вероятности пропуска самозванца и отказа законному пользователю. По результатам тестирования на одной и той же базе данных, регулярно проводимого в Национальном институте стандартов и технологий США (NIST), эта ошибка находится в диапазоне 3 – 5%, так что суммарная ошибка равна удвоенной величине, т.е. 6 – 10%.

Введение

Настоящий обзор посвящен, в основном, результатам работ в области распознавания дикторов за последнюю декаду. Более ранние результаты обобщены в монографиях [88, 169, 213], а также в обзорах (9, 18, 28, 56, 105, 141, 170, 190]. Знакомство с предысторией предмета исследований необходимо не столько для эрудиции исследователя, сколько для понимания того, что уже было сделано, какие направления оказались тупиковыми, а какие – перспективными. Это дает возможность осознано сконцентрироваться на наиболее обещающих направлениях и, даже в случае обманутых надежд потери драгоценного времени будут минимизированы. К сожалению, в мировой литературе наблюдаются многочисленные случаи повторения уже известных (некоторым специалистам) результатов, которые авторы выдают за пионерские откровения. Поэтому трудно переоценить роль аналитических обзоров не только для начинающих исследователей, но и для специалистов, поскольку число работ в области распознавания диктора исчисляется многими сотнями, а время для поиска источников ограничено, да и сами источники могут оказаться труднодоступными или находиться в таких изданиях, которые трудно заподозрить в интересе к данному предмету. Вероятно, качественный обзор вносит не меньший вклад, чем исследование, критически повлиявшее на развитие этой области науки.

Исторически, исследование проблемы распознавания диктора начиналось в интересах юриспруденции, криминалистики и контр-разведки. Поэтому сначала проблема рассматривалась как аналогичная проблеме распознавания отпечатков пальцев, чему свидетельство известная работа

[102], которая ввела в обращение вводящий в заблуждение термин "отпечатки голоса" (voice prints). Это направление исследовало контурные линии равного уровня энергии на сонограммах, вызывающие ассоциации с папиллярными узорами пальцев. Достаточно быстро выяснилось, что узоры контурных линий подвержены многочисленным видам изменчивости, и никак не могут использоваться в качестве признаков для распознавания диктора.

В связи с развитием информационных технологий в настоящее время в распознавании диктора, помимо государственных учреждений, заинтересованы бизнес-структуры и многочисленные категории пользователей информационных услуг. Несмотря на интенсивные научные исследования и появляющиеся время от времени сообщения о феноменальной эффективности разработанных систем распознавания, реальное применение, за исключением узких областей, сильно ограничено, что подтверждается регулярными годовыми отчетами Gartner Group, констатирующих, что лишь около 1% процента объема потенциальных пользователей удовлетворено эффективностью коммерческих систем распознавания диктора.

Нельзя сказать, что прогресс в этой области вообще отсутствует. Периодические испытания на фиксированных базах данных, организуемые Национальным институтом стандартов и технологии США (NIST), демонстрируют постепенное повышение эффективности систем распознавания диктора. Тем не менее, практические применения этих систем немногочисленны, поскольку их характеристики еще далеки от требований потребителей. Поэтому и возникает необходимость в том, чтобы время от времени выполнить анализ состояния дел в этой области с тем, чтобы определить наиболее перспективные направления.

Области применения

Распознавание диктора подразделяется на два направления: идентификацию и верификацию. При верификации пользователь предъявляет в том или ином виде свой идентификатор, и система распознавания должна подтвердить или отвергнуть этот идентификатор. При этом в большинстве случаев пользователь заинтересован в подтверждении его идентификатора, и старается не вносить в речевой пароль вариаций, которые отсутствовали в период обучения на его голос.

При идентификации диктор не указывает своего идентификатора, и система распознавания должна установить, принадлежит ли речевой сигнал голосу одного из дикторов, прошедших обучение. Важный случай идентификации состоит в определении того, принадлежат ли два речевых сигнала голосу одного и того же диктора. Эти два вида идентификации существенно отличаются по условиям принятия решений. Поэтому необходимо ввести дополнительную классификацию: идентификацию с обучением и идентификацию без обучения. Эти виды идентификации могут пересекаться, если установлении принадлежности имеется достаточная база представителей речевых сигналов целевого диктора.

Постановка задачи распознавания диктора зависит от конкретной цели. Подробнее условия верификации и идентификации будут рассматриваться ниже.

Криминалистика и судебная экспертиза

Необходимость в определении того, принадлежит ли голос подозреваемого записям речи в телефонных каналах возникает при анализе телефонных звонков в случае ложных сообщений, наркодеятельности, вымогательства или сексуальных домогательств. При этом, в отличие от верификации, предметом анализа могут быть лишь записи речевых сигналов, подлежащих сравнению, либо вновь выполненные записи речи подозреваемого. В последнем случае подозреваемый обычно не заинтересован в его идентификации, и его речь может быть сознательно искажена. К тому же, условия такой записи, выполненной, например, в тихой комнате для допросов, могут сильно отличаться от условий, в которых подлежащие сравнению речевые сигналы были сгенерированы и переданы по каналу связи, а записанные фразы могут быть разными. В криминалистике подозреваемого могут попросить прочитать текст, соответствующий транскрипции ранее записанной речи, но, как показал опыт, этот прием не очень эффективен.

Представители органов криминалистики заинтересованы в том, чтобы получить однозначный ответ от принадлежности биометрических параметров. Например, исследовательская группа Федерального бюро расследований США утверждает, что в отношении отпечатков пальцев приемлемо только однозначное решение – "совпадает/не совпадает", и не должны использоваться никакие оценки типа "возможно, вероятно, может быть" [192]. Но даже и в отношении отпечатков пальцев такая позиция мало обоснована. Считается, что вероятность ложного совпадения отпечатков пальцев порядка

10^{-6}, хотя на этот счет отсутствуют статистически достоверные исследования. Что же касается автоматического распознавания отпечатков пальцев, то вероятность ложного опознания гораздо выше – около 2% для 4 пальцев (Fingerprint Verification Competition, 2004). Не случайно при верификации личности в важных случаях требуются отпечатки всех десяти пальцев. Решение об идентичности только по одному отпечатку вообще имеет высокий риск ошибки.

Научные основы применения технологии идентификации голоса в криминалистике обсуждались в [21, 22, 34, 39, 56, 60]. Общее мнение состоит в том, что идентификация по голосу отличается от отпечатков пальцев и генетике, где вариации очень малы, и нет абсолютно надежного метода для определения того, принадлежат ли речевые сигналы одному и тому же человеку. В криминалистике распознавание диктора может иметь только вероятностный характер, т.е. с указанием правдоподобия того, что два речевых сигнала принадлежат одному и тому же человеку. В условиях телефонного канала проблематично даже распознавание пола или возраста. В силу малой выборки речевых сигналов доверительный интервал оценки правдоподобия принадлежности двух записей речи одному и тому же диктору столь велик, что однозначное решение невозможно.

Специальный тест с парным сравнением речевых сигналов длительностью 5 с показал 53% правильного распознавания фонетистами, которым было разрешено пользоваться любыми техническими средствами, и 46% - не фонетистами. В других тестах диапазон составлял 38 -76% [21]. Эти оценки наглядно показывают степень неопределенности принятия решений.

В соответствии с этим мнением, в судебной практике США, Великобритании и Франции экспертное заключение об идентичности записей речи не принимается в качестве юридического доказательства. Это вполне логично, поскольку в практике уголовного расследования при визуальной идентификации личности требуется сравнение с некоторым количеством похожих лиц, тогда как решение об идентичности голосов, основанное только на сравнении перехваченных записей речевого сигнала и голоса подозреваемого, без сравнения с голосами множества других дикторов, содержит высокий риск ошибки. Этот риск может не остановить от принятия решения в некоторых случаях, как это было описано в книге А.И.Солженицына "В круге первом", но обязанность научного сообщества состоит в том, чтобы предупредить об отсутствии оснований для категорических решений.

Радио-разведка, контр-разведка, антитеррористический мониторинг

В этих областях идентификация диктора не носит юридического характера. Поэтому решение о степени близости голосов принимается на основе вероятностных количественных оценок, и само по себе не является двузначным. В такой постановке государственных организаций задача идентификации голосов имеет определенную специфику, связанную искажениями и помехами в каналах связи. Поскольку фонетическое содержание сравниваемых речевых сигналов обычно различно, то государственные организации заинтересованы в исследованиях распознавания диктора независимо от контекста.

Конечно, и в этом случае разработка методов идентификации голосов содержит негативный аспект, связанный со злоупотреблениями виде вмешательства в частную жизнь или надзор за оппозицией правящего режима.

Безопасность

Государственные организации, бизнес-структуры и частные лица заинтересованы в обеспечении безопасности использования современных информационных технологий. Обман и злоупотребления со стороны своих сотрудников наносят ущерб около 6% годовой прибыли, составляя, в среднем, около $100.000 на каждый случай (в 14.6% случаев потери превысили $1.000.000) (Association of Certified Fraud Examiners, 2004). В банковской сфере потери от злонамеренной деятельности сотрудников финансовые потери могут достигать огромных величин. Несанкционированный доступ к конфиденциальной информации о финансовой деятельности компании, контрактах и планах чреват не только потерями, но и полным банкротством.

Передаваемые по телефону параметры кредитной карты в 12% случаев подслушиваются с последующим воровством денег с карты (American Bankers Association). Аналогично, параметры кредитной карты перехватываются в системах электронной торговли или в банкоматах. Украденные суммы исчисляются сотнями миллионов долларов в год.

Существует ряд ситуаций, в которых человеку необходимо подтвердить свое право на распоряжение материальными или информационными ресурсами, доступ к информации или в помещение, сейф и т.д. Подтверждение такого права осуществляется с помощью документов

(паспорта, удостоверения личности, пропуска), физических (ключи, кредитные карты) или электронных средств (коды авторизации, пароли). В ряде случаев такие средства верификации личности либо неудобны, либо не обеспечивают необходимой степени защиты. Согласно решению *Federal Financial Institution Examination Council, USA,* от 2005 года, использование однофакторной методологии аутентификации личности (т.е. подтверждения личности с помощью ПИН-кода или буквенно-цифрового пароля) является неадекватным средством защиты в системах удаленного доступа к финансам. Поэтому, в дополнение к таким традиционным средствам, целесообразно использовать биометрические параметры человека. Преимущество биометрии заключается в том, что эти параметры всегда находятся при человеке, их нельзя забыть, потерять, передать другому человеку, украсть и довольно трудно воспроизвести.

Принципиальный недостаток всех методов биометрии, кроме речевого, состоит в постоянстве используемого биометрического кода, т.к. отпечатки пальцев или ладоней, рисунок радужной оболочки и черты лица неизменны для индивидуума. Этот недостаток препятствует применению этих методов в случаях, требующих особо высокой надежности идентификации личности, поскольку неизменный биометрический код может быть считан путем злонамеренного вторжения в программу распознавания.

В отличие от биометрии по фиксированным параметрам, верификация по голосу обладает практически неограниченным потенциалом для снижения ошибки за счет использования все более длинных речевых сообщений. Верификации по голосу может использоваться в темноте, на расстоянии, в частности, по стандартному телефонному каналу, в условиях, когда невозможно получить изображение лица.

Примеры конкретных применений верификации диктора охватывают широкий спектр приложений:

- распоряжение финансовыми процессами по электронным или телефонным каналам (управление банковским счетом, электронная коммерция, подтверждение права пользования кредитной картой);
- разрешение на смену пароля или PIN-кода;
- доступ к компьютеру или отдельным программам компьютера (вход в Интернет, доступ к конфиденциальным документам, базам данных и т.д.);
- разрешение на вход в помещение, открывание сейфа;
- управление механизмами и системами (например, запуск двигателя автомобиля);
- мониторинг того, кто, когда и к каким компьютерным ресурсам имел доступ.

Добавление акустического распознавания диктора в несколько раз уменьшает ошибку распознавания по лицу/фигуре, но добавление визуальной информации лишь ненамного улучшает решение по акустике [47].

В определенных ситуациях, например, при получении команд пилотом, необходимо убедиться в том, что команда отдана лицом, имеющим на это право. Голос человека, передающего команду, может быть не знаком получателю информации, и в этом случае полезна автоматическая идентификация группы лиц, уполномоченных на отдание команд. Очевидно, такая ситуация может существовать не только в авиации.

Некоторые заболевания коры правого полушария головного мозга могут привести к потере способности к распознаванию голоса [193, 194]. Такое заболевание может быть достаточно скрытым, и в определенных условиях автоматическая идентификация голоса становится необходимой.

Сегментация дикторов

Сегментация дикторов в потоке разговора разных дикторов (audio-indexing, diarization) необходима при разметке звуковых стенограмм, теле-конференций, радио- и теле-передач, интервью, расшифровке записей разговоров на вечеринке (cocktail-party), видео-клипы каникул [11, 41, 46, 67, 72, 139, 151, 167, 190, 200].

Извлечение мета-данных в виде пола говорящего, предмета дискуссии, имен участников позволяет осуществить автоматический поиск и индексирование. При сегментации, так же, как и при криминалистической экспертизе, диктор должен рассматриваться как не желающий сотрудничать, поскольку, в отличие от верификации, у нет задачи быть распознанным.

Различают методы сегментации, при которых определяются только моменты смены дикторов (speaker turn detection), и методы, в которых распознается диктор (speaker clustering). По данным [67], ошибка EER распознавания диктора составляет 15.4%.

В определенных условиях для сегментации достаточно распознать пол диктора. Если доступен достаточно длительный сегмент речевого высказывания, то распознавание пола может быть выполнено практически безошибочно [168]. На коротких сегментах типа ударного гласного ошибка правильного распознавания мужского пола составляет 5.3%, а женского пола 3.1% [181].

Удобство

В современном обществе человек вынужден запоминать пароли и PIN-коды для обеспечения доступа к разнообразным услугам. Эти данные часто теряются или забываются, что создает досадные проблемы и требует восстановления или смены этих кодов. Согласно оценкам Meta Group, каждый клиент, в среднем, звонит в службу помощи клиентам примерно 15 раз в год, причем от 20% до 50% звонков содержит просьбы о возобновлении или смене пароля (отчеты Gartner Group). Каждый акт возобновления пароля требует общения с человеком-оператором, и занимает, в среднем, около 3 мин, если пользователь помнит все правильные ответы на вопросы, задаваемые с целью подтверждения его права на возобновление пароля (типа "назовите девичью фамилию Вашей матери"). Это время может быть и гораздо больше.

В случае необходимости использования удаленного доступа, например, по телефону, удобство голосовой верификация пользователя приобретает решающее значение. Круглосуточный, ежедневный доступ, например, к управлению банковским счетом или финансовыми операциями обеспечивает оперативность и удобство при активной деловой деятельности.

Доступ к информации может осуществляться с помощью речевого общения и без формального процесса верификации. Например, если заранее известно, что речевой запрос на получение информации доступен только определенному лицу, то при получении такого запроса по умолчанию предполагается, что он принадлежит этому лицу, и выполняется оценка вероятности вторжения самозванца, на основании которой и принимается решение о доступе.

Экономия эксплуатационных расходов

Автоматическая верификация пользователя позволяет исключить участие человеческого персонала в процессе санкционирования, повышая степень защищенности системы, экономя время и зарплату персонала при ежедневной и круглосуточной работоспособности. Информационно-справочные службы или службы помощи клиентам получают значительную экономию средств от исключения человека- оператора и режима 7*24 (семь дней в неделю, 24 часа в сутки) готовности обслуживания. Согласно Gartner Group, смена пароля обходится при человеческом обслуживании от \$10 до \$31 (в среднем, \$25) на каждый случай, что на каждую 1000 клиентов экономит до \$375.000 в год.

Финансовая выгода также может состоять в предотвращении финансовых или иных потерь, в сравнении с которыми затраты на обеспечение безопасности доступа более, чем оправданы.

Показатели эффективности систем верификации

В число таких показателей входят ошибки первого (вероятность пропуска самозванца) и второго рода (вероятность отказа), вероятность отказа от обучения, взломоустойчивость, реакция на заболевания и алкоголь, действия при многократном отказе, задержка принятия решения.

Вероятность пропуска самозванца указывает на степень защиты от злонамеренного вторжения, тогда как вероятность отказа законному пользователю определяет удобство эксплуатации системы распознавания. В зависимости от темперамента и условий применения разные люди по разному реагируют на отказ. Поэтому при некоторой вероятности отказа, независимо от надежности системы относительно злонамеренного вторжения, пользователь сам откажется от эксплуатации такой системы. Критический уровень вероятности отказа считается 10%, хотя на этот счет не известно достоверных исследований.

Согласно статистической теории решений, соотношение между ошибками первого и второго рода зависит от порога принятия решений, которые, в свою, очередь, определяются различными факторами, в том числе и индивидуальными предпочтениями пользователя. В [58] в качестве интегральной оценки эффективности системы распознавания диктора рассматривается средне-геометрическое $E_{сг} = \sqrt{E_{пропуска} E_{отказа}}$). Однако средне-геометрическое - не постоянная величина, и она увеличивается с уменьшением вероятности пропуска до 1 – 2%, делая более предпочтительным отказ перед признанием диктора.

Другая оценка – взвешенная сумма стоимости отказа и пропуска. Это превосходная оценка для реальных систем. В [17, 83] интегральная оценка включает априорные вероятности появления самозванца и риск:

$$E_{дикт} = C_{проп} P_{дикт} P_{проп} + C_{отк} P_{сам} P_{отк},$$

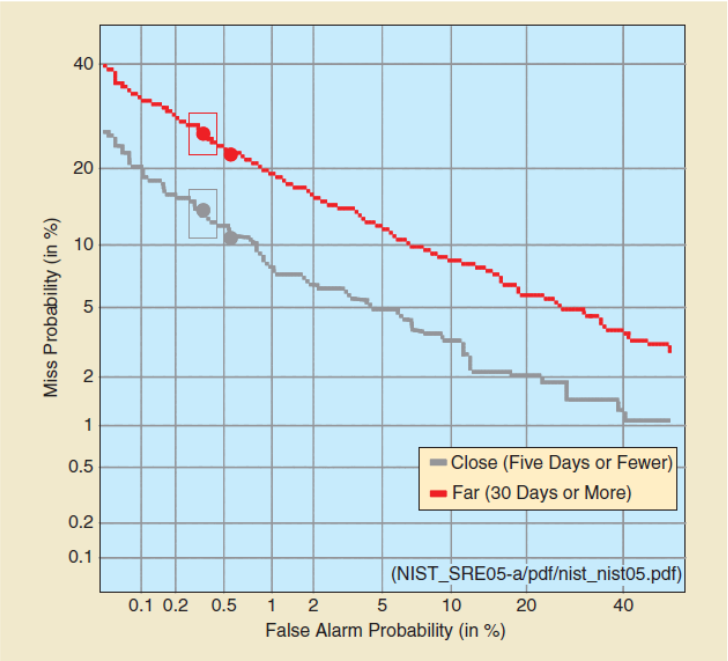
где $P_{дикт}, P_{сам}$ - вероятности появления диктора и самозванца, $P_{проп}, P_{отк}$ - вероятности пропуска самозванца и отказа целевому диктору, $C_{проп}, C_{отк}$ - риск (стоимость) пропуска самозванца и отказа целевому диктору.

В этой оценке вероятности появления самозванца и риск должны устанавливаться самим пользователем, что лишает возможности сравнения различных систем распознавания. К тому же, совершенно неясно, на каком основании должны устанавливаться количественные показатели $P_{дикт}, P_{сам}$ и $C_{проп}, C_{отк}$.

Национальный институт стандартов и технологий США (NIST) использует более простую оценку в виде функции минимальной стоимости детектирования ошибки detection cost function (DCF)

$$DCF_{дикт} = 0.1 P_{проп} + 0.99 P_{отк}$$

Наиболее полную характеристику системы распознавания дает функция зависимости вероятности пропуска самозванца от вероятности отказа (DET – Detection Error Trade-off) [136] . Обычно эта зависимость представляется в логарифмическом масштабе для удобства визуального анализа и сравнения различных методов, как это показано на Рис.1 из [58] при обсуждении разницы в показателях системы при распознавании вскоре после обучения и через 1 месяц:



[FIG1] Effect of time between enrollment and test recordings, NIST-SRE '05.

Рис. 1. Зависимость вероятности пропуска самозванца от вероятности отказа целевому диктору по [58]. Логарифмический масштаб.

Наиболее популярна оценка в виде равной вероятности пропуска и отказа (ERR – Equal Error Rate). На Рис.1 ERR равна 5% при распознавании вскоре после обучения, и около 8% при более позднем распознавании. Оценка ошибки как ERR не является полной оценкой характеристик системы распознавания, т.к. основана на произвольном установлении порога принятия решений, но она дает примерное представление о качестве системы, и позволяет сравнивать различные системы. Поэтому ERR можно использовать лишь только при сравнении различных систем, но она не пригодна для оценки конкретной системы. Необходимо также учитывать, что суммарная ошибки – пропуска и отказа, равна удвоенной величине ERR.

Как видно на Рис. 1, попытка уменьшить вероятность пропуска самозванца приводит к экспоненциальному росту вероятности отказа. Например, если потребовать, чтобы вероятность

пропуска была равна 1%, вероятность отказа становится равной 40%, что абсолютно неприемлемо для пользователя.

Для того, чтобы была уверенность в достоверности указанных разработчиком вероятностей ошибок, необходимо оценивать статистическую значимость как по объему выборки речевых сигналов при тестировании, так и по числу тестированных дикторов. При малых объемах доверительный интервал может оказаться таким большим, что декларируемые вероятности ошибок не имеют ничего общего с действительностью. В [58] предлагается эвристическое правило 30, вытекающее из биномиального распределения: для того, чтобы быть уверенным на 90%, что полученная оценка верна, должно наблюдаться, по крайней мере, 30 ошибок. Вероятность пропуска в 1% означает, что должно быть не меньше 3000 испытаний для самозванца, а 0.1% отказа - не меньше 30000 испытаний для истинного диктора. Правда, при этом неясна справедливость предположения о независимости испытаний, на основании которой получено это правило.

В дополнение, должны быть указано, совпадали ли условия обучения и тестирования, а если нет - то насколько ухудшаются оценки. Как правило, коммерческие системы распознавания не сообщают всю необходимую информацию, и, как будет показано ниже, независимое тестирование показывает, что реальные характеристики в несколько раз хуже объявленных.

Средняя по множеству тестированных дикторов оценка вероятностей первого и второго рода также не вполне описывает эффективность системы распознавания. Голоса большинства дикторов обладает умеренной способностью к подтверждению личности. Такие дикторы, по [58], называются овцами. Голоса других дикторов легко имитируются, и их называют ягнятами. Дикторы, голоса которых часто принимаются за голоса других дикторов, называются волками. Наконец, дикторы с нестабильными параметрами голоса и плохим показателем распознаваемости, называются козлами. Характеристика системы распознавания должна включать в себя процентное соотношение всех типов голосов, принимавших участие в тестировании.

Эффективность системы верификации диктора зависит от того, насколько она уязвима для злонамеренного вторжения самозванца с помощью имитации голоса целевого диктора или воспроизведения заранее записанной речи, а также попытки использования родственников с похожими голосами. У пользователей вызывает опасение отказ от верификации в случае простудных заболеваний. Необходимо также предусмотреть такие действия в случае многократного отказа от верификации, которые не увеличивают риск злонамеренного вторжения.

Установлено, что на результат распознавания диктора по голосу влияют уровень образования и интеллект (хотя эти факторы неудобно обсуждать по этическим соображениям).

Вторжение в систему верификации

С самого начала разработок систем верификации возник вопрос о возможности имитации голоса целевого диктора. Эксперименты с профессиональными имитаторами показали, что успех такой подделки голоса невелик [170], в частности, как показали исследования France Telecom на большой базе данных, вероятность пропуска имитатора в среднем не превышает вероятности пропуска самозванца из базы данных. Имитаторы лучше всего подделывают интонационный контур и карикатурно подчеркивают явные особенности речи, но глубинные факторы, определяющие индивидуальные свойства речи, имитировать не удается. Голоса родственников одного и того же пола обладают определенным сходством, но даже голоса близнецов, имеющих сходную анатомию речевого тракта, могут различаться по динамике управления [34, 149].

В последнее время разрабатываются методы трансформации одного голоса в другой, и это создает опасность для систем распознавания диктора [22, 157]. Так, в [22] сообщается, что формирование голоса самозванца с помощью преобразования к параметрам пользователя увеличивает ошибку верификации до 50%. Этот новый фактор заставляет пересмотреть распространенные методы анализа речи, и требует использование таких параметров, которые трудно воспроизвести при трансформации голосов.

Подмена голоса целевого диктора может выполняться с помощью записанных скрытным микрофоном или перехваченных в канале связи речевых сигналов. Такое вторжение особенно опасно для систем верификации с фиксированным паролем. Один из способов проверки факта вторжения с помощью записанных речевых сигналов в системах с фиксированным словарем состоит в сравнении двух одинаковых слов или фраз в системах с фиксированным паролем или там, где произнесение диктора управляется самой системой верификации. Если обнаруживается полная идентичность произнесений, то это может свидетельствовать о вторжении. Правда, если такой способ обнаружения

вторжения известен злоумышленнику, то он может быть преодолен неоднократной записью одних и тех же слов и использованием другой записи при повторном предъявлении.

Существуют, однако, физические основания для уменьшения опасности вторжения с помощью воспроизведения подслушанных сигналов при входе в систему верификации через микрофон. Один из этих факторов заключается в отличии характеристик и положения скрытного микрофона от характеристик и положения относительно диктора микрофона, через который осуществляется штатный вход в систему верификации. Другой фактор состоит в искажениях амплитудно-частотной характеристики сигнала, создаваемой акустическими свойствами воспроизводящей системы. Искажения речевого сигнала, создаваемые приемниками звука и каналами связи, обычно рассматриваются как вредные факторы, ухудшающие эффективность систем верификации. Однако в случае попытки подмены речевого сигнала эти факторы могут оказаться полезными, увеличивая вероятность отказа вследствие несовпадения параметров голоса целевого диктора, сформированных в процессе обучения, и параметров сигнала вторжения.

Если речевой сигнал или его параметры были перехватываются в канале связи, и затем вводятся в этот канал с целью вторжения, то противодействие такому вторжению может заключаться в использовании специального кода, встраиваемого в речевой сигнал (стеганографии, "водяные знаки") [63, 64, 65, 93]. Этот код должен меняться от одной передачи сигнала к другой. В частности, это может быть метка момента времени передачи, которая сравнивается текущим временем на приемном конце. Контрольная информация вводится, например, в коэффициенты линейного предсказания таким образом, что ее удаление разрушает речевой сигнал, а если обнаруживается контрольный сигнал от предыдущей записи, то фиксируется вторжение. Конечно, если известен алгоритм кодирования и декодирования контрольного сигнала, то он может быть подделан – его нужно хорошо засекречивать. Вместе с тем, эти метки не должны мешать распознаванию сигналов от целевого диктора.

Необходимо понимать, что никакой способ борьбы со злонамеренным вторжением дает 100% гарантии, и лишь затрудняет попытки взлома. Например, если в системе верификации злоумышленнику оказывается доступен тот блок, в котором принимается решение, то достаточно подменить код правильного распознавания.

Может показаться, что вероятность пропуска самозванца может быть сведена к нулю путем сдвига порога решения в сторону вероятностного распределения признаков речи целевого диктора. Это не так. Обычно распределения целевого диктора и дикторов из базы пересекаются во всем диапазоне параметров, так что при любом смещении порога вероятность пропуска самозванца не может стать нулевой. Кроме того, такой сдвиг порога приводит к непропорционально быстрому возрастанию вероятности отказа целевому диктору, и это приводит уже к отказу пользователей эксплуатировать систему верификации.

Субъективная эффективность распознавания диктора

При разработке систем автоматического распознавания диктора желательно знать потенциально минимальную вероятность ошибки, и использовать ее в качестве показателя эффективности конкретной системы распознавания. Такой потенциальной вероятностью может служить вероятность ошибки, полученная в экспериментах с распознаванием голосов людьми.

В этих экспериментах выяснилась роль длительности сигнала, подлежащего распознаванию. Конкретные оценки варьируются от исследования к исследованию, но общая тенденция такова: чем длительнее речевой сигнал, тем с большей вероятностью распознается голос диктора. В ранних экспериментах [213] было установлено, что по мере предъявления все более длительных сегментов речи вероятность правильного распознавания возрастает почти вдвое: от 56% для отдельных гласных, до 98% для предложений, и около 87% для двусложных слов. В других экспериментах оценки оказались значительно ниже: 31% для слова "hello", 66% - для фразы и 83% для 30 с речи. Эти ранние результаты затем многократно воспроизводились в последующих работах вплоть до последнего времени. Как упоминалось выше, специальный тест с парным сравнением речевых сигналов длительностью 5 с показал 53% правильного распознавания фонетистами, которым было разрешено пользоваться любыми техническими средствами, и 46% - не фонетистами. В других тестах диапазон оценок составил 38 - 76% [21]. Вероятность ошибки идентификации в экспериментах [80], где требовалось определить, принадлежат ли два предложения одному и тому же диктору, в среднем по дикторам, оказалась близкой к 22%, т.е. около 78% правильной идентификации.

Вероятность правильного распознавания зависит и от условий эксперимента [169]. Если сравниваемые речевые сигналы были записаны в одних и тех же условиях, то, по [172], вероятность

правильной идентификации фраз составляет около 92%, а если сравниваемые сигналы были записаны через разные каналы, то вероятность правильной идентификации 86%, т.е. ошибка увеличивается почти вдвое.

Индивидуальность голоса определяется анатомией речевого тракта, характеристиками источника голосового возбуждения, системой управления артикуляцией. Поэтому следует ожидать разницы в субъективной вероятности правильного распознавания диктора в разных контекстах. Возможно, этим и определяются разногласия в результатах тестирования с участием аудиторов.

На узнаваемость голоса виляет и то, говорит ли он на родном или не родном языке. В первом случае вероятность правильного распознавания составляет около 95% (ошибка 5%), тогда как во втором – 87% (ошибка 13%, т.е. в два с лишним раза выше) [201].

Форма речевого тракта лучше всего проявляется при нейтральном положении артикуляторных органов. Система управления артикуляцией учитывает особенности анатомии тракта, адаптируя артикуляцию с тем, чтобы акустические параметры речевого сигнала попали в диапазон, характерный для данного языка. Поэтому, чем больше деформируется форма речевого тракта, тем меньший вклад в акустические характеристики вносит анатомия тракта. Ближе всего к нейтральному состоянию соответствует артикуляция гласного /э/. По данным [213], ошибка распознавания диктора по этому гласному составляет 10%, она возрастает до 14 – 17% для гласных /*а, о, и*/. Хуже всего дикторы распознаются для гласного /*у*/ с ошибкой до 40%. В экспериментах [120] средняя ошибка распознавания диктора, произносящего изолированный гласный /*а*/, составила около 50%.

Влияние голосового источника проявляется в том, что звонкие фрикативные /*з, ж*/ обеспечивают меньшую ошибку распознавания (21 – 26%), чем соответствующие им глухие фрикативные /*с, ш*/ (56 – 63%) и аффрикаты /*ч, ц*/ (46 – 50%) [213]. Форма носовых полостей должна сказываться на характеристиках назальных звуков. Ошибка распознавания диктора по назальному /*м*/ составляет около 38% [213].

В экспериментах с обратным проигрыванием речи установлено, что разборчивость разрушается почти полностью, но голос диктора все же остается до некоторой степени узнаваемым [193, 194], хотя ошибка распознавания при этом довольно велика – около 55% [213]. Это как-будто свидетельствует об относительно малой роли контекста. Вместе с тем, в [213] было найдено, что семантическое содержание речевого сигнала (изолированные звуки или слоги, бессмысленная последовательность раздельных слогов, осмысленные слова, фразы) существенно влияет на узнаваемость голоса с распределением ошибок от 45% для изолированных гласных до 10% для фраз. Конечно, здесь также влияет и длительность речевого сигнала.

В криминалистике иногда требуется указать возраст диктора. Перцептивное восприятие возраста по телефону характеризуется, в среднем, примерно 80% правильных ответов [23, 38, 147]. При этом возраст молодых (18-24 года) людей занижается (64% точных оценок), а пожилых (60-66 лет) – завышается (78% точных оценок). Наиболее точно оценивается возраст в интервале 46 – 52 лет (96% точных оценок). Наиболее точно оценивается принадлежность возраста к одной из трех групп: молодых, среднего возраста и пожилых.

Оценки роста и веса противоречивы по голосу диктора: в [119] сообщается о хорошей их различимости, тогда как в некоторых источниках утверждается обратное.

Изменчивость

Параметры речевого сигнала для одного и того же произнесения варьируются как в силу нестабильности произнесения самим диктором (intra-speaker variability), так и вследствие разнообразия внешних условий. К внутренним факторам изменчивости относятся стиль, темп и громкость речи, а также речь на фоне шума. Внешние факторы включают в себя вид и уровень помех в акустическом и электронном канале связи, искажение речевого сигнала приемниками звука и реверберацией помещения. Внешние факторы в виде диалектных особенностей сказываются на условиях формирования референтной базы данных дикторов, с которыми выполняется сравнение при вычислении меры принадлежности голоса целевому диктору.

Помимо обычного, разговорного, стиля речи используется речь с повышенной (hyper articulation) или весьма невнятной артикуляцией (hypo articulation). Стиль речи проявляется в акустических характеристиках речевого сигнала [185] и темпе речи. Темп речи зависит также от длительности высказывания, сложности обсуждаемого предмета, настроения [91] и эмоционального состояния (48). Известно, что разные люди имеют исходно разный темп речи [180, 191]. Темп зависит от возраста (быстрее всего говорят дикторы в возрасте около 40 лет), пола (мужчины говорят

быстрее женщин независимо от длительности фразы), географического происхождения даже в одной и той же стране, при чтении темп ниже, чем при спонтанной речи [57, 91]. Эти факторы также необходимо учитывать при формировании референтной базы дикторов.

Изменение громкости речи диктора приводит к изменении амплитудно-частотных характеристик речевого сигнала. В частности, известен так называемый эффект Ломбарда, состоящий в повышении уровня высокочастотных компонент речевого сигнала при непроизвольном повышении громкости в присутствии помех.

Характеристики голоса диктора подвержены непрерывному изменению во времени, поэтому разница во времени, когда выполнялось обучение, и времени акта распознавания может существенно повлиять на показатели системы верификации. Если распознавание выполняется через несколько недель после обучения, то ошибка удваивается [117]. На голос также влияют состояние здоровья, например, ларингит, заболевания легких.

Изменчивость амплитудно-частотных характеристик речевого сигнала связана и с различием типов микрофонов, расстояния от диктора до микрофона и направления микрофона. Близко расположенные микрофоны улучшают отношение «речевой сигнал - акустические шумы среды», однако при этом возникает эффект ближнего акустического поля, при котором амплитудно-частотные характеристики сигнала в низкочастотной области сильно зависят от расстояния до микрофона. К тому же, использование головных гарнитур с близко расположенным микрофоном неприемлемо для большинства пользователей.

Различные положения мобильного телефона: щеки-плечо, ухо-плечо, далеко от рта, с сигаретой во рту, в ладонях (что создает дополнительный резонанс в области 2 кГц и выше 4 кГц), приводят к тому, что формантные частоты сдвигаются, пропадают или появляются ложные форманты [27, 113, 174].

Реверберация помещения приводит к искажению амплитудно-частотных характеристик речевого сигнала [181], а также к длительному затуханию колебаний на формантных частотах звука, предшествующего смычке. Реверберация также порождает ложные пики в сигнале-остатке [206].

Компенсация канала

По мнению авторо [162], успех в распознавании диктора в гораздо большей степени зависит от метода компенсации канала, чем от выбора признаков. Существование проблемы канала было осознано в конце 90-х годов [154, 163, 164, 203].

Рассматривается несколько способов компенсации характеристик каналов: model-based, score-based, feature-based. В model-based методе, использующем модель канала применяют либо стерео-запись для многих типов микрофонов [142] с последующим вычислением преобразования между ними, либо распознавание типа микрофона в режиме дикторо-независимого распознавания. Применяется также вычитание среднего кепстра [8, 74]. Логарифмирование спектра или кепстра переводит влияние канала из мультипликативной помехи в аддитивную, что позволяет использовать методы спектрального или кепстрального вычитания. Компенсация канала по RASTA-PLP [85, 133] опирается на предположение о стационарности характеристик канала, что позволяет отфильтровать сигналы с модуляциями от 1 до 16 Гц.

В score-based (HNORM) методе определяется тип микрофона и при обучении и при распознавании, причем вычисляется функция одного нормального распределения для различия в модели диктора между разными микрофонами, и компенсация выполняется путем вычитания смещения, зависящего от микрофона, и масштабирования по среднему и дисперсии. Если обучение происходит для разных микрофонов, то вариант нормализации состоит в том, что параметры нормализации вычисляются для каждого канала в отдельности, и при верификации определяется тип канала путем выбора наибольшего правдоподобия, хотя это, конечно, связано с ошибками

В feature-based методе не требуется стерео-записи и ручной разметки на типы микрофона ни в обучении, ни в распознавании, но нелинейное преобразование применяется к признакам, таким, так лог-спектр или кепстр [25].

Компенсация характеристик канала (в частности, среднее значение кепстра), связанная с различием расстояния до микрофона в [198] выполнялась путем оценки расстояния до 4 микрофонов, расположенных на плоскости Т-образно, по времени прихода сигнала. Если расстояние до микрофона меняется динамически, то возникает проблема адаптации в реальном времени. Более эффективный метод нормализации, пригодный для верификации, но не для идентификации, описан в [10], где распределение признаков приводится к нормальному распределению.

Эффективный способ нормализации каналов предложен в [97, 99, 100, 101]. Различие между каналами моделируется в явном виде с помощью совместного факторного анализа joint factor analysis (JFA). Параметры каждого канала представляются векторами "собственных каналов" (eigenchannels), которые находятся по большой базе данных.

Классификация задач распознавания диктора

Выше упоминались две основные задачи – это верификация и идентификация диктора. В системах санкционирования доступа идентификация диктора может применяться в тех случаях, когда, по условиям эксплуатации, предусматривается проверка личности только одного диктора. В этом случае нет необходимости в указании идентификатора этого диктора. Решение принимается путем сравнения голоса на входе системы идентификации с характеристиками единственного пользователя. Вариант задачи идентификации состоит в идентификации диктора из некоторой группы, а не отдельного диктора (например, при групповом допуске в помещение или доступе к информации). Такая постановка задачи рассматривается в [171]. При проверке принадлежности голоса диктора к группе аккредитованных пользователей, также не нужно указывать идентификатор диктора.

Если в группу дикторов входят как мужчины, так и женщины, то для сокращения перебора сначала выполняется распознавание пола. В зависимости от длительности речевого сигнала и метода распознавания, ошибка распознавания пола может быть близкой к нулю [168], либо достигать величин 5 – 6% [181].

Если число потенциальных пользователей не слишком велико, то идентификация конкретного диктора также возможна без использования идентификатора. В этом случае поочередно выполняется проверка гипотезы о принадлежности поступившего на вход систем идентификации голоса к каждому из представленных в группе дикторов. Идентификация осуществляется путем выбора того диктора, для которого достигнуто наилучшее значение правдоподобия. Число дикторов, для которых еще возможно использование идентификации вместо верификации, определяется вычислительной мощностью системы и допустимым временем задержки. В случае массового обслуживания, когда число пользователей слишком велико, идентификация практически неприменима.

Верификация диктора может рассматриваться как задача дихотомии: один против всех. Однако конкретные алгоритмы сопоставления параметров целевого диктора с параметрами дикторов в референтной базе могут быть разными. Сравнение параметров может происходить с объединенными параметрами дикторов из референтной базы, поочередно с каждым из дикторов из этой базы, либо с параметрами типичных представителей дикторов из этой базы, найденных путем кластеризации (eigen-voices). Множество самозванцев может быть известно или нет. Это множество устанавливается, если распознавание диктора выполняется среди фиксированной группы.

Ввод идентификатора диктора в системах верификации может осуществляться различными способами. Если пользователь обращается к системе верификации через компьютер, то наиболее простой и надежный способ состоит в использовании алфавитно-цифрового кода или выбора соответствующего идентификатора в меню. Такой способ не снижает устойчивость системы верификации к попыткам злонамеренного вторжения, поскольку идентификатор всего лишь указывает на область значений параметров голоса, с которыми и сравнивается поступивший на вход речевой сигнал.

В качестве идентификатора могут использоваться отпечатки пальцев или изображение радужной оболочки глаза. Эти биометрические параметры могут быть представлены в виде кода, который и служит идентификатором. Поскольку такие системы характеризуются ненулевой ошибкой распознавания, то порог принятия решений должен быть установлен таким образом, чтобы минимизировать вероятность отказа от распознавания. Применение таких способов существенно усложняет систему верификации диктора.

При удаленном доступе, например, по телефонному каналу, в качестве идентификатора может служить номер мобильного телефона, с которого осуществляется запрос. При использовании кабельного телефона общего пользования идентификатор может быть задан с помощью клавиатуры телефона.

В тех случаях, когда пользователю доступен только микрофон, идентификатор может определяться с помощью системы распознавания речи или путем применения специального устройства, генерирующего последовательность звуковых импульсов. Эта последовательность импульсов должна формироваться в виде уникального кода.

Относительно вида речевых высказываний, на основе которых решается задача распознавания диктора, различают методы, зависящие от текста, и независимые от него. Пространство признаков, в котором выполняется распознавание диктора в большинстве известных систем одно и то же, и не зависит ни от контекста, ни от языка. Анализ таких признаков приводится ниже.

В криминалистике речевые данные произвольны, и поэтому исследование в интересах такого применения сосредоточены на методах распознавания, независимых от контекста. Перенос такого подхода на задачи санкционирования доступа представляется мало перспективным. Считается, что диктору при каждом акте распознавания удобно произносить любые фразы. На самом деле, это требует от диктора каждый раз сознательно формировать новый текст, что создает определенную когнитивную нагрузку. Поэтому в действительности в таких системах пользователи обычно произносят одну и ту же фразу [117]. Это превращает систему распознавания, формально не зависящую от контекста, в систему с фиксированным паролем, обладающую наименьшей устойчивостью к вторжению самозванца с помощью воспроизведения подслушанного и записанного пароля.

Недостаток системы фиксированным паролем состоит еще в том, что каждого диктора пользователя такой пароль произволен, а референтная база дикторов формируется с использованием другого множества высказываний. Такая разница в речевом материале ухудшает эффективность системы распознавания, тогда как создание референтной базы специально для конкретного пароля практически неосуществимо.

Оптимальный компромисс между удобством пользователя и эффективностью системы распознавания состоит в использовании фиксированного словаря, состоящего из небольшого количества слов, хорошо знакомых любому диктору. Такой словарь, например, может состоять из числительных от 0 до 9 [84, 117, 150]. Пароль, состоящий из последовательности таких слов должен случайно изменяться при каждом акте распознавания [182]. Таким образом избегается опасность вторжения с помощью записанного пароля, свойственная системам с фиксированным паролем. По данным [117], специально подобранный контекст может снизить ошибку в 2 раза по сравнению со словарем числительных.

Технологически системы распознавания разделяются на системы индивидуального и коллективного пользования. При санкционировании, например, доступа к операционной системе или каким-либо данным в персональном компьютере, распознавание диктора выполняется непосредственно в этом компьютере. При удаленном доступе, например, по телефонному каналу или Интернету, распознавание может осуществляться на сервере с множественным доступом.

Анализ речи, признаки

Индивидуальность акустических характеристик голоса определяется тремя факторами: механикой колебаний голосовых складок, анатомией речевого тракта и системой управления артикуляцией. В спонтанной речи также проявляются индивидуальные особенности использования словаря и оборотов речи.

Размеры голосовых складок, масса, жесткость и вязкие свойства складок, давление в легких находятся в основе процессов автоколебаний складок. Частота колебаний складок и форма импульсов объемной скорости потока, протекающего через голосовую щель, влияют на форму огибающей спектра речевого сигнала и его временные параметры. Геометрические размеры различных отделов речевого тракта и боковые полости (грушевидные полости в области гортани, две носовые полости, гайморовы полости), а также механические свойства тканей речевого тракта определяют его резонансные частоты и скорость затухания колебаний на резонансных частотах. В спектре речевого сигнала это проявляется как частоты и ширина его пиков.

Система управления артикуляцией формирует просодические характеристики: динамику частоты основного тона, длительность фонетических сегментов, скорость движения артикуляторов, а также эффекты коартикуляции, которые по разному проявляются у разных дикторов. Например, наибольшее влияние индивидуальности найдено для /*u*/ [86]. Коартикуляция для назализованных звуков содержит информацию о дикторе [184].

Индивидуальность стиля речи проявляется на достаточно длительных высказываниях, и может быть полезна, например, в задачах сегментации дикторов в потоке речевых сигналов, содержащих речь нескольких дикторов. Акустически стиль реализуется в виде контура частоты основного тона, длительности слов и его сегментов, ритмики ударных сегментов, длительности пауз, уровня громкости [116].

Признаки, связанные с диалектными особенностями, рассматривались в [57, 94] особенности произношения рассматривались в [5, 30, 31, 92, 146, 211], а просодические характеристики - в [1, 2, 40, 157, 173, 177, 200]. Особенности стиля исследуются и на артикуляторном уровне [125], а также с применением метода мульти-язычного моделирования [110, 166]. Особенности диалекта не распределены по всем произнесениям, а проявляются в отдельных фонемах, а длительность и частота появления пауз могут характеризовать иностранца [202].

Пространство признаков, в котором принимается решение о личности диктора, должно формироваться с учетом всех факторов процесса речеобразования: голосового источника, резонансных частот речевого тракта и их затуханий, а также динамикой управления артикуляцией. В частности, рассматриваются следующие параметры голосовго источника: средняя частота основного тона, контур частоты основного тона, флюктуации частоты основного тона и форма импульса возбуждения. Спектральные характеристики речевого тракта описываются огибающей спектра и его средним наклоном, формантными частотами и их полосами, долговременным спектром или кепстром (см. ниже). Кроме того, рассматриваются также длительность слов, ритм (распределение ударений), уровень сигнала и частота и длительность пауз [116]. В [182] в качестве признаков использовались частота основного тона, три формантных частоты на переходных и стационарных участках гласных, параметры огибающей спектра фрикативных, а также общая длительность слова и относительные длительности сегментов речи.

Считается, что просодические характеристики и признаки высокого уровня более устойчивы, но обладают меньшей различающей способностью. К тому же они легче всего поддаются имитации [7, 108].

Выделенные параметры могут использоваться в виде временных рядов или в виде долговременных оценок, предложенных в [66,73], и активно используемых в системах верификации, независимых от контекста. При этом необходимо подвергать анализу достаточно длительные отрезки речевого сигнала. В [79] было установлено, что вплоть до длительности в 18 сек долговременный спектр зависит от контекста, а в [123] считают, что для достижения приемлемых результатов типичная длительность фазы обучения должна быть не менее 5 мин, хотя в отдельных приложениях она может быть равна и 10 сек.

В [138] сообщается, что наиболее важный фактор индивидуальности голоса – это частота основного тона F_0, за ней следуют формантные частоты, размер флюктуаций F_0 и наклон спектра. В [173] высказывается мнение, что признаки, связанные с F_0, обеспечивают наилучшую разделимость голосов, а за ними следуют энергия сигнала и длительность сегментов. Логарифмическое представление F_0 более информативно, чем сама частота основного тона [103, 176]. Среднее значение частоты основного тона в долговременной статистике исследовалось в [36, 103, 135, 148, 176, 177], а дисперсия и скос распределения - в [12, 36, 103, 118].

В другой работе наиболее важным фактором считаются формантные частоты [120]. В частности, четвертая форманта практически не зависит от типа фонемы и характеризует тракт [187]. Это вытекает из свойств управляемости резонансными частотами речевого тракта, рассмотренных в [180]. Механика артикуляции такова, что в области высоких частот на сужение в речевом тракте приходятся как пучность, так и узел соответствующих собственных функций акустических колебаний, и это не позволяет управлять частотами высших резонансов.

В работах по распознаванию диктора доминирует метод кепстрального преобразования спектра речевых сигналов (метод впервые предложен в [52]). Схема этого метода такова: на интервале времени в 10 – 20 мс вычисляется текущий спектр мощности, а затем применяется обратное преобразование Фурье от логарифма этого спектра (кепстр) [54, 89], и находятся коэффициенты кепстра:

$$c_n = \frac{1}{\Theta}\int_0^{\Theta} \log\left|S(j\omega,t)\right|^2 e^{-jn\Omega\omega}\,d\omega$$

$\Omega = 2\pi/\Theta$, Θ - верхняя частота в спектре речевого сигнала, $\left|S(j\omega,t)\right|^2$ - спектр мощности.. Число кепстральных коэффициентов n зависит от требуемого сглаживания спектра, и находится в пределах от 20 до 40.

Если используется гребенка полосовых фильтров, то коэффициенты дискретного кепстрального преобразования вычисляются как

$$c_n = \sum_{m=1}^{M} [\log Y(m)] \cos[\frac{\pi n}{M}(m-\frac{1}{2})],$$

где $Y(m)$ – выходной сигнал m-го фильтра, c_n – n-й коэффициент кепстра.

Свойства слуха учитываются путем нелинейного преобразования шкалы частот, обычно в шкале *мел*. Эта шкала формируется исходя из присутствия в слухе так называемых критических полос, таких, что сигналы любой частоты в пределах критической полосы неразличимы. Шкала *мел* вычисляется как

$$M(f) = 1125 \cdot \ln(1 + f/700),$$

где f - частота в Гц, M – частота в мелах.

Существует и другая, похожая шкала *барк*, такая, что разность между двумя частотами, равная критической полосе, равна 1 *барк*. Частота B в барках вычисляется как

$$B = 13 arctg(0.00076 f) + 3.5 arctg\left(\frac{f}{7500}\right)$$

Коэффициенты кепстрального преобразования формируют пространство, в котором и производится распознавание диктора. Эти коэффициенты сокращенно обозначаются как MFCC – Mel Frequiency Cepstral Coefficients. Число используемых коэффициентов от 10 до 30. Часто используются первые и вторые разности по времени кепстральных коэффициентов, что втрое увеличивает размерность пространства принятия решений, но улучшает эффективность распознавания диктора [211].

Кепстр описывает форму огибающей спектра сигнала, в которой интегрируются характеристики источников возбуждения (голосового, турбулентного и импульсного) и формы речевого тракта. В экспериментах по субъективному распознаванию голоса было установлено, что огибающая спектра сильно влияет на узнаваемость голоса [90]. Поэтому использование того или иного способа анализа огибающей спектра в целях распознавания диктора оправдано.

Вместо вычисления спектра речевого сигнала с использованием дискретного преобразования Фурье на коротком интервале времени, используется также амплитудно-частотная характеристика сигнала, найденная по коэффициентам линейного предсказания речи [89]. Например, в [171] 10 коэффициентов линейного предсказания пересчитываются в 12 коэффициентов кепстра. В этом случае шкала частот вычисленного спектра – линейная, что приводит к проигрышу по сравнению с нелинейной шкалой. Недостаток такого метода заключается и в том, что наиболее отлаженные процедуры линейного предсказания представляют передаточную функцию речевого тракта в виде дробно-рациональной функции, содержащей только полюса.

Однако в речевом тракте присутствуют и разветвления – грушевидные полости в области гортани, носовые и гайморовы полости. Эти разветвления создают нули в передаточной функции речевого тракта, существенно влияющие на вид спектра. Морфология речевого тракта не меняется в процессе речеобразования и лучше всего характеризует индивидуальность диктора. Частотные области, связанные с боковыми полостями мало маскируются контекстом. Информация о диктора неравномерно распределена по частотам.

Даже для высокого не-назализованного /*u*/ и звонкой смычке сильная связь с носовой полостью происходит через вибрации небной занавески [50, 186]. Параназальные полости также оказывают сильное влияние на акустику, что проявляется при насморке, когда связь с этими полостями перекрыта. Грушевидные полости создают анти-резонансы между 4 и 5 кГц, что является важным признаком для распознавания диктора [49, 51].

Спектр речевого сигнала может вычисляться и с помощью гребенки полосовых фильтров, так или иначе описывающих свойства периферического отдела системы слухового восприятия у человека. Наиболее популярна модель, описанная в [153], в которой спектрально-временные характеристики речевого сигнала анализируются гребенкой фильтров, называемых gammatone. Весовая функция каждого фильтра есть

$$g(t) = t^{n-1} e^{-bt} \cos(\omega t + \varphi)$$

где n – порядок функции (обычно n=4), b – определяет ширину полосы пропускания, ω - центральная круговая частота ,а φ - фазовая константа, которая обычно принимается равной нулю. Преобразование Лапласа для такого фильтра есть [124]

$$G(s) = \frac{6(-b^4 - 4b^3 s - 6b^2 s^2 - 4bs^3 - s^4 + 6b^2\omega^2 + 12bs\omega^2 + 6s^2\omega^2 - \omega^4)}{(b^2 + 2bs + s^2 + \omega^4)^4}$$

В системе gammatone шкала частот может быть выбрана произвольно, что создает гибкость в разработке методов анализа речи. Логарифмическая шкала частот обеспечивает относительно большую устойчивость кепстральных коэффициентов по сравнению с использованием преобразования Фурье в качестве первого этапа анализа.

MFCC изначально исследовалось в интересах распознавания речи с подавлением индивидуальных характеристик диктора. Этот метод обеспечивает хорошую разрешающую способность в низкочастотной области и низкую в высокочастотной, что полезно для анализа фонетических характеристик, но плохо для распознавания диктора. С этой целью вместо мел-шкалы применялось монотонное преобразование шкалы частот [140]. Неравномерное преобразование исследовалось в [128]. В этой работе было найдено три информативные области: 100-300 Гц (влияние голосового источника), 4-5 кГц (грушевидные полости) и 6.5 – 7.8 кГц – (возможно, влияние согласных). Небольшая область – в районе 1 кГц. В соответствии с зонами наибольшей чувствительности было выполнено нелинейное преобразование спектра, а затем вычислялись 32 кепстральных коэффициента. Сообщается о снижении ошибки распознавания на 20%. Результаты работы [128] подтверждают ранее найденные в [14] информативные области в спектра речевых сигналов. В этой работе была найдено, что наибольшая информация о дикторе находится в полосах ниже 600 Гц и выше 3000 Гц. Частотный диапазон стандартного телефонного канала 300 -3400 Гц обрезает высокие частоты, ухудшая различимость голоса диктора.

Анализ параметров голосового источника обычно выполняется на сигнале-остатке в линейном предсказании [158]. Согласно предположению Г.Фанта о независимости источника и речевого тракта, сигнал остаток рассматривается как аналог импульса возбуждения акустических колебаний в речевом тракте, т.е. как производная от объемной скорости воздушного потока через голосовую щель. В аудиторских экспериментах было установлено, что прослушивание сигнал-остатка дает достаточную информацию для субъективного распознавания диктора [69].

В [77, 143, 188] распознавание диктора выполняется в пространстве кепстральных коэффициентов, вычисленных по спектру сигнала-остатка. Вероятность ошибки распознавания EER диктора при использовании только кепстральных параметров сигнала-остатка в [158] весьма высока (от 28% до 64%), тогда как в [77] приводится оценка в 5%. В [188] сообщается, что добавление этих параметров к MFCC спектра речевого сигнала снижает ошибку EER с 5.7% (суммарная ошибка 11.4%) до 4% (суммарная ошибка 8%). В этих работах также отмечается значительное увеличение ошибки, если распознавание выполняется через определенное время после обучения. Различные эмоции также влияют на параметры голосового источника [3, 159].

В [59] на множестве звонких сегментов речи диктора вычисляется первая собственная функция сигнала-остатка в низкочастотной области (eigen-residual) и средняя огибающая по Гильберту шумовой компоненты в высокочастотной области. Около 1000 звонких сегментов достаточно для надежной оценки обеих компонент. Распознавание диктора по этим компонентам оказалось весьма надежным. В Табл. 1 представлены результаты оценки EER для распознавания на базах данных разного объема. Как видно из этой Таблицы, увеличение количества дикторов более чем в 3 раза приводит к увеличению ошибки почти в 2 раза, но наименьшая ошибка сопоставима с результатами распознавания на основе кепстральных коэффициентов.

Таблица. 1. Ошибка распознавания диктора (%) по собственной функции сигнала-остатка и огибающей шумовой компоненты.

	168 дикторов	630 дикторов
Собственная функция	5.88	11.43
Огибающая	8.76	17.14
Обе компоненты	1.98	3.65

Физика голосообразования такова, что от одного периода к другому меняется и длительность периода и амплитуда возбуждения. Эти факторы обозначаются как jitter и shimmer. Вариации периода основного тона (jitter) у здоровых людей находятся в диапазоне 0.1 – 1%. У людей с некоторыми заболеваниями гортани этот диапазон значительно шире, и перспективен для диагностики. Однако, для того, чтобы обеспечить необходимую точность анализа jitter (хотя бы 10%) необходимо использовать частоту дискретизации речевого сигнала от 100 кГц до 1 мГц, что неприемлемо в системах общего пользования. Тем не менее, считается, что микро-вариации основного тона могут быть надежно определены после низкочастотной фильтрации в полосе до 1000

Гц. В [76] оценки jitter и shimmer выполняются путем совместной оценки параметров модели голосового источника Фанта-Лилиенкрантца (LF-model).

Косвенные оценки параметров голосового источника получаются путем анализа амплитудных соотношений в определенных областях речевого спектра. Такой анализ особенно часто применяется для распознавания пола диктора или в задачах диагностики патологии гортани. В число этих признаков входят, например, разность амплитуд первых двух гармоник основного тона (H_1-H_2), разность амплитуды первой гармоники и амплитуды третьей форманты (H_1-A_3), а также разности амплитуд второй и четвертой гармоники основного тона (H_2-H_4) [44]. Кроме того, в качестве признака используется отношение уровней шумовой и гармонической компоненты в речевом сигнале.

Распознавание пола диктора в [181] было выполнено в пространстве параметров модели голосового источника, найденных путем решения обратной задачи – от сигнала-остатка к модели производной от объемной скорости потока через голосовую щель, и, далее, к модели динамики площади голосовой щели. Было получено снижение ошибки распознавания пола на 40% по сравнению с решением, основанным только на частоте основного тона. Есть основания полагать, что такой способ перспективен и для распознавания диктора, особенно в системах, независимых от контекста.

Решающие правила

В силу того, что в подавляющем большинстве систем распознавания диктора используется одно и то же пространство признаков в идее кепстральных коэффициентов, их первых и вторых разностей, основное внимание уделяется построению решающих правил. Наиболее популярны метод аппроксимации плотности вероятности в пространстве признаков взвешенной смесью нормальных распределений (GMM – Gauss Mixture Models), метод опорных векторов (SVM – Support Vector Machines), метод скрытых Марковских моделей (HMM – Hidden Markov Models), искусственные нейронные сети, а также модификации факторного анализа.

Метод GMM непосредственно вытекает из теоремы, гласящей, что любая функция плотности вероятности может быть представлена как взвешенная сумма нормальных распределений:

$$p(x) = \sum_{j=1}^{k} w_j \varphi(x; \theta_j),$$

$$\sum_{j=1}^{k} w_j = 1,$$

где $\varphi(x; \theta_j)$ – функция распределения многомерного аргумента x с параметрами θ_j,

$$\varphi(x; \theta_j) \equiv p(x \mid \mu_j, R_j) = \frac{1}{(2\pi)^{n/2} |R_j|^{1/2}} e^{-\frac{1}{2}(x-\mu_j)^T R_j^{-1}(x-\mu_j)}, \; x \in \mathbb{R}^n$$

w_j – её вес, k – количество компонент в смеси. Здесь n – размерность пространства признаков, $\mu_j \in \mathbb{R}^n$ – вектор математического ожидания j-й компоненты смеси, $R_j \in \mathbb{R}^{n \times n}$– ковариационная матрица.

Применение метода GMM оправдывается двумя факторами. Первый фактор состоит в необходимости описания плотности вероятности в многомерном пространстве признаков, сформированном для референтной базы дикторов [161, 162]. В [165] указывается, что для построения адекватной GMM для референтной базы дикторов необходимо, чтобы в ней содержались речевые сигналы длительностью в десятки и даже сотни часов. Второй фактор связан с аппроксимацией плотности вероятности для целевого диктора, особенно в системах, независимых от контекста.

Для оценивания параметров смеси $\Theta_k = (w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$ используется метод максимума правдоподобия или метод максимума апостериорной вероятности. Обычно задается фиксированное число компонент смеси – от 256 до 2048, а главные оси компонент направлены вдоль координатных осей пространства признаков (Рис. 2). Это связано с большим числом вычислений, для сокращения которых используется диагональная матрица ковариаций.

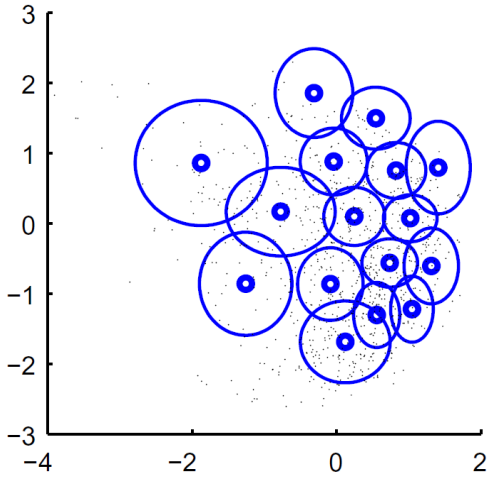


Рис. 2. Аппроксимация распределения смесью нормальных распределений.

Если обучающая выборка содержит мало векторов параметров, то распределение вероятностей выгоднее описывать только одним нормальным распределением. В этом случае используется полная матрица ковариаций [14, 15, 17, 28, 212]. Модель с полной матрицей ковариаций в общем случае описана в [209]. Для малых обучающих выборок матрица ковариаций GMM распределений, даже диагональная, плохо обусловлена. Выход состоит в ограничении дисперсии по минимальному значению, т.е. если в диагонали дисперсия $\sigma^2 < \sigma^2_{\min}$, то $\sigma^2 = \sigma^2_{\min}$. Это необходимо выполнять адаптивно, для каждого параметра отдельно и индивидуально для каждого диктора [16].

Метод GMM может рассматриваться как расширение метода векторного квантования [162, 165]. При векторном квантовании, также известном как метод центроидов, создается кодовая книга для непересекающихся областей в пространстве признаков, обычно с помощью кластеризации методом K-means [126]. Векторное квантование является простейшей моделью в системах распознавания диктора независимо от контекста [26, 81, 96, 104, 178, 179]. В отличие от векторного квантования, GMM использует перекрывающиеся области в пространстве признаков.

С целью уменьшения сложности оптимизационной задачи поиска параметров смеси нормальных распределений применяется алгоритм EM (expectation-maximization) [19], в котором алгоритм кластеризации K-means может использоваться для поиска начальных приближений. Рассматриваются алгоритмы, в которых требуется небольшое число или даже отсутствие итераций в алгоритме EM [106, 111, 155]. При формировании алгоритма EM важную роль играют критерии эффективности и разделимости.

Критерий эффективности $C(\Theta_k)$ описания выборки смесью из k компонент, включает в себя штраф на число компонент. Один из таких критериев описан в [129]:

$$C(\Theta_k) = -2L(\Theta_k) + 2E(\Theta_k) + \nu(\Theta_k)\log m\,.$$

Здесь $L(\Theta_k)$ – логарифм функции правдоподобия, $E(\Theta_k)$ – энтропия, $\nu(\Theta_k)$ – число свободных параметров в смеси Θ_k, m – число элементов в выборке. Логарифм функции правдоподобия $L(\Theta_k)$ определяется как

$$L(\Theta_k) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j \varphi(x_i; \theta_j)\,.$$

Энтропия $E(\Theta_k)$ записывается формулой

$$E(\Theta_k) = -\sum_{i=1}^{m}\sum_{j=1}^{k} g_{ij} \log(g_{ij}) \geq 0\,.$$

Здесь $g_{ij} \equiv \mathrm{P}(\theta_j \mid x_i)$ – апостериорная вероятность того, что обучающий объект x_i был сгенерирован

j-й компонентой смеси.

Критерий разделимости $S(j;\Theta_k)$, характеризующий качество описания j-й компонентой смеси принадлежащих ей объектов, описан в [145]:

$$S(j;\Theta_k) = \int f_j(x;\Theta_k) \ln \frac{f_j(x;\Theta_k)}{\varphi(x;\theta_j)} dx,$$

где $f_j(x;\Theta_k)$ – локальная плотность выборки для j-го распределения. Распределение с наибольшим $S(j;\Theta_k)$ имеет наихудшую оценку локальной плотности и, следовательно, является первым кандидатом на разделение.

В [204] все гауссовы компоненты референтной модели предварительно кластеризуются в дерево, и соответствующие акустические пространства отображаются в структурно разделенные области. При этом достигается ускорение примерно в 16 раз с ухудшением ошибки лишь на 1%.

Поскольку GMM не опирается на фонетические элементы, то в системах, независимых от контекста, возникает разногласие между обучающим и распознаваемым контекстом. Этот недостаток пытаются преодолеть либо с помощью фонетического дерева решений [43, 82] или создания GMM для каждого фонетического элемента (phonetic GMM (PGMM)) [37, 61, 78, 152] или части слога [20].

Метод опорных векторов (SVM) активно используется в различных системах распознавания образов после публикации монографии [195]. Этот метод позволяет построить гиперплоскость в многомерном пространстве, разделяющую два класса, например, параметров целевого диктора и параметров дикторов из референтной базы. Гиперплоскость вычисляется с использованием не всех векторов параметров, а только специально выбранных. Эти вектора и называются опорными. Поскольку разделяющая поверхность в исходном пространстве параметров не обязательно соответствует гиперплоскости, то выполняется нелинейное преобразование пространства измеренных параметров в некоторое пространство признаков более высокой размерности. Это нелинейное преобразование должно удовлетворять требованию линейной разделимости в новом пространстве признаков. Если это условие выполняется, то разделяющая поверхность в гиперплоскости строится методом опорных векторов. Очевидно, что успех применения метода опорных векторов зависит от того, насколько удачно подобрано нелинейное преобразование в каждом конкретном случае при распознавании дикторов.

Метод опорных векторов применяется для верификации дикторов часто в комбинации с методом GMM (29, 32, 33, 35, 132, 197] или HMM [115]. Метод опорных векторов применяется также к просодическим параметрам [68, 173] и признакам высокого уровня [30].

К распознаванию дикторов применяется и метод скрытых Марковских моделей (HMM), хорошо зарекомендовавший себя в задачах автоматического распознавания речи [13, 16, 23, 42, 70, 114, 137, 144, 171, 175, 199]. В частности предполагается, что для коротких фраз длительностью в несколько секунд для контестно-зависимого подхода лучше всего применять фонемно-зависимые HMM, а не модели на основе вероятностей перехода от кадра к кадру длительностью 10 – 20 мс [211]. Метод скрытых марковсикх моделей может использоваться в совокупности с методом GMM [75, 114].

Искусственные нейронные сети (ANN – Artificial Neuron Net) применялись при распознавании диктора в [62, 83, 160, 206]. В [159] ANN применялась к сигнал-остатку на периоде основного тона.

Эффективность применения конкретного типа классификатора зависит от того, происходит ли сравнение параметров целевого диктора с общим распределением параметров в референтной базе, с N-ближайшими по результатам обучения, с каждым диктором из референтной базы, или с дикторами-кластерами этой базы.

Кластеризация (eigenvoices)

Несоответствие между объемом обучающей выборки и размерностью пространства признаков создает серьезные трудности при формировании решающих правил и ставит под сомнение любые оценки эффективности распознавания диктора. Размерность пространства признаков измеряется десятками. Так, число кепстральных коэффициентов может варьироваться от 10 до 30. Обычно к ним добавляются первые и вторые разности по времени, так что размерность пространства признаков для каждого кадра находится в диапазоне от 30 до 90. В общем случае это число должно быть умножено

на количество кадров в речевом высказывании, что приводит к размерностям в сотни параметров. Длительность обучения не может быть слишком большой по чисто психологическим причинам. Поэтому и число векторов параметров в обучающей выборке несопоставимо мало по сравнению с тем числом, которое должно быть для того, чтобы оценки вероятностного распределения в пространстве параметров были состоятельными.

Проблему недостаточности обучающих данных для каждого диктора пытаются решить, используя референтную базу данных. С этой целью в референтной базе находятся типичные представители голосов (eigenvoices, anchor speakers), и модель диктора формируется как взвешенная сумма вероятностных распределений параметров типичных голосов с критерием максимума правдоподобия (ML – Maximum Likelihood) [131] или максимума апостериорной вероятности (MAP – Maximum Aposterior Probability) [6, 98, 112, 130, 131, 189]. Для очень коротких обучающих выборок альтернатива критерию максимума апостериорной вероятности состоит в применении критерия линейной регрессии максимального правдоподобия (MLLR – Maximum likelihood linear regression), [95, 107, 122, 130, 131, 134, 183]. В этом методе модель диктора находится как нелинейная функция от собственных векторов в пространстве MLLR.

Супер-векторы

Еще один популярный метод перевода данных разной размерности в единственный вектор, соответствующий произнесению – это создание так называемого супер-вектора. GMM диктора может рассматриваться как супер-вектор [35, 53, 121, 165]. Компонентами такого супер-вектора являются значения математических ожиданий смеси GMM. Например, вектор измерений размерностью *d* для *k* компонент гауссовой смеси представляется как единственный вектор размерностью *dk* [32, 35]. Этот супер-вектор может использоваться как входные данные для SVM.

Супер-вектора позволяют компенсировать вариации произнесения от сессии к сессии [25, 101, 196]. Любая вариация рассматривается как влияние среды, микрофона или контекста, и считается вредной для распознавания. Один из методов нормализации – факторный анализ, позволяющий использовать GMM [97], где гауссовский супер-вектор рассматривается как линейная комбинация компонент, зависящих от диктора и от канала, которые считаются статистически независимыми. Такая нормализация позволяет восстанавливать условия, отсутствующие при обучении.

При таком подходе возникают два вопроса – как создать супер-вектор произнесения, и как оценить и применить компенсацию вариабельности сессий в пространстве супер-вектора. В применении к SVM супер-вектор создается как обобщенная последовательность линейных дискриминантов путем проекции в пространство ядра SVM с использованием полиномиального разложения [29, 35]. В последней работе, например, ядро гауссовского супер-вектора создается путем ограничения меры невязки Кульбака-Лейблера между гауссовыми смесями. А в [208] супер-вектор создается путем ограничения расстояния Бхаттачарайя. Принципиальная разница между супер-векторами MLLR и гауссовскими супер-векторами состоит в используемой модели речи – фонетической (HMM) и гауссовской, а также в методе адаптации MLLR и максимума апостериорной вероятности.

Агрегирование классификаторов

Каждый метод принятия решений (классификатор) обладает определенными преимуществами и недостатками, и по-разному реагирует на различие в условиях обучения и распознавания, а также на особенности голоса разных дикторов. Поэтому возникает желание так использовать решения разных классификаторов, чтобы достичь минимально возможной ошибки распознавания. Существует обширная литература по этому вопросу, посвященная, в основном, математическим аспектам проблемы. Основной прием состоит в том, чтобы учесть качество каждого классификатора, и решение принимается как взвешенная по этим оценкам сумма решений [45, 71, 87, 109]. Исходя из теории доказательств Демпстера-Шэйфера, при агрегировании должны участвовать решения не меньше, чем трех классификаторов. В применении к распознаванию дикторов агрегирование рассматривалось в [4], где каждый классификатор характеровался четырьмя факторами: матрицей ошибок для каждого диктора, списком дикторов с плохим решением, списком дикторов с правильным решением, списком дикторов (соседей) для каждого произнесения. Сообщается о снижении ошибок распознавания на 3 - 25% по сравнению с минимальной ошибкой, достигаемой при использовании любого одного классификатора.

Один из методов усиления простых классификаторов, основанный на комбинировании примитивных "слабых" классификаторов в один "сильный" называется бустинг (boosting). Под

"силой" классификатора подразумевается эффективность (качество) решения задачи классификации, которое обычно измеряется средним числом ошибок классификации на обучающей выборке.

Строгий алгоритм машинного обучения для произвольных (ε, δ) при обучении на достаточно большой случайной выборке *S* с вероятностью $1-\delta$ выдает гипотезу классификации h_S, которая имеет ошибку обобщения не более ε. Кроме этого, время работы такого алгоритма должно полиномиальным образом зависеть от $1/\varepsilon$, $1/\delta$ и размера выборки *S*. Слабый алгоритм машинного обучения по определению должен удовлетворять тем же свойствам, за исключением того, что то же самое выполнено для хотя бы одного $\varepsilon < \frac{1}{2} - \gamma$, где $\gamma > 0$ -- константа.

Алгоритм бустинга строит сильный алгоритм машинного обучения по слабому алгоритму машинного обучения путем многократного прохождения по обучающей выборке и увеличения веса примеров, на которых слабый алгоритм дает большую ошибку обучения.

Наиболее известным алгоритмом бустинга является алгоритм AdaBoost [71].

Сравнительный подход к задаче предсказания

Еще один подход заключается в сравнении результатов прогноза для разных методов. Правильный прогноз или правильное решение ведут к меньшим потерям, чем неправильные. При традиционном статистическом подходе оцениваются потери прогноза в сравнении с некоторой идеальной моделью принятия правильных решений, которая обычно основана на некоторой статистической модели, описывающей наблюдаемые данные. При этом сначала оцениваются параметры статистической модели на основе наблюдений, а потом производится прогноз на основе этой модели при оцененных параметрах.

При сравнительном подходе вместо одной идеальной модели рассматривается набор возможных моделей, которые называются экспертными стратегиями, или просто, экспертами. Множество таких экспертных стратегий может быть конечным или бесконечным и даже несчетным. Используя исходы, поступающие в режиме онлайн, экспертные стратегии производят прогнозы будущего исхода. Прогнозирующий алгоритм может наблюдать прогнозы экспертных стратегий и оценивать их эффективность в прошлом. После этого алгоритм делает свой прогноз. Прогнозы этого алгоритма оцениваются в сравнении с прогнозами экспертных алгоритмов. Обычно производится сравнение потерь выбранного алгоритма за некоторый период прогнозирования с потерями наилучшего на ретроспективе эксперта. Сравнение может производиться как в наихудшем случае, так и в среднем, если алгоритм использует рандомизацию. Заметим, что распределение вероятностей, которое использует рандомизированный алгоритм, является внутренним вспомогательным распределением алгоритма; оно не имеет никакого отношения к источнику, генерирующему исходы.

В качестве основы для определения функции потерь могут использоваться многие количественные методы оценки качества классификации или предсказания. Например, в случае задачи классификации потери экспертного метода – это просто сумма ошибок за время обучения. В случае вероятностного метода вычисления прогноза используются абсолютная, квадратичная, логарифмическая функции потерь. В последнем случае, функция потерь совпадает с логарифмом функции правдоподобия $L\left(\Theta_k\right)$, определенной выше.

Некоторые алгоритмы смешивания экспертных стратегий эффективно работают со специальными функциями потерь – логарифмической и квадратичной. Они имеют меньшую ошибку предсказания. Другие алгоритмы рассчитаны на произвольные функции потерь. В этом случае, в качестве функции потерь может использоваться плотность

$$p\left(x\right) = \sum_{j=1}^{k} w_j \varphi\left(x;\theta_j\right),$$

а также другие функции общего вида.

Алгоритм взвешенного большинства

Простейший алгоритм на точное предсказание будущего исхода называется алгоритмом взвешенного большинства. Этот алгоритм обучается в режиме онлайн. Для простоты, предполагается, что имеется два возможных исхода 0 и 1. Имеются N экспертов (стратегий), которые на каждом шаге t выдают предсказания P_t^N=0 или 1. Изучающий алгоритм обозревает в режиме

онлайн бинарную последовательность $\omega_1\omega_2\omega_3...\omega_{t-1}$ и прогнозы каждого из экспертов $p_1^1....p_t^1$, …. $p_1^N....p_t^N$ на всех шагах, включая шаг *t*, и предсказывает будущий исход p_t=0 или 1. Классический алгоритм взвешенного большинства, был предложен Литтлстоуном и Вармутом в 1989г. [127].

Неполное описание

Нередко встречаются ситуации, когда проще отказаться от использования какого-то признака, чем рисковать увеличением ошибки вследствие его ненадежности. В теории кодирования этот прием называется стиранием. Преимущество такого приема заключается в том, что место ненадежно определенного признака точно известно. В результате стирания на вход классификатора подаются векторы с неполным составом компонент. Принципы распознавания в таких ситуациях рассматриваются в подходе, который называется missing data approach [55]. Распознавание можно выполнять в подпространстве, образованном отключением стертого признака. Например, если частота какой-то из трех формант не принята к распознаванию, то решение может выполняться в двумерном пространстве, образованном надежно определенными формантами.

Другой подход заключается в попытках использования целостного образа. Аналогия этой ситуации встречается в физиологии под названием константности восприятия или гештальта. Стертый признак может заменяться, например, его средним значением по всей обучающей выборке диктора [208]. Еще один способ состоит в подстановке значения признака, принадлежащего полномерному вектору, ближайшему к вектору со стертым признаком в подпространстве этого вектора.

Заключение

Идентификация диктора по произвольному тексту применяется в криминалистике для установления принадлежности разных речевых высказываний одному и тому же диктору, при сегментации записей стенограмм или интервью на участки речи, принадлежащие каждому из участников разговора, а также при установлении личности без указания его идентификатора среди сравнительно небольшого множества дикторов. Верификация диктора выполняется с использованием его идентификатора, который может быть предъявлен в любой форме – алфавитно-цифрового пароля (PIN кода), электронной карты-идентификатора, или даже фразы, специфичной для данного диктора, например, его имени, отчества и фамилии. Вероятность принять речь другого диктора за голос целевого диктора и вероятность отказа целевому диктору при верификации может быть значительно ниже, чем при идентификации.

В качестве входных параметров, в пространстве которых выполняется распознавание диктора, могут использоваться частота основного тона и его вариации, формантные частоты, длительности сегментов речевого высказывания, в том числе паузы. При анализе длительных речевых сигналов, как, например, при сегментации дикторов, специфическим для диктора может оказаться используемый им лексикон. Выделение этих параметров из речевого сигнала требует разработки сложных алгоритмов, а погрешности, например, в определении формантных частот, могут оказаться довольно велики. Поэтому наибольшее распространение получили параметры в виде коэффициентов кепстра, который вычисляется по огибающей спектра, полученного через преобразование Фурье, с помощью гребенки фильтров, либо по передаточной функции речевого тракта, найденной методом линейного предсказания. В дополнение к коэффициентам кепстра используются также их первые и вторые разности по времени.

Несовпадение характеристик канала при обучении и распознавании негативно влияет на эффективность систем распознавания. Поэтому рассматриваются различные способы компенсации (нормализации) характеристик канала.

Значительные усилия прилагаются к разработке математических методов принятия решений. Среди них наибольшее распространение получили метод аппроксимации плотности вероятности взвешенной суммой нормальных распределений, скрытые Марковские модели, метод опорных векторов и искусственные нейронные сети.

С целью стандартизации оценок различных систем распознавания дикторов был введен критерий равной ошибки EER, который указывает на вероятность ошибки при условии равенства вероятностей пропуска самозванца и отказа целевому диктору. К настоящему времени наилучшие исследовательские системы распознавания диктора характеризуются величинами EER порядка 3 – 5%. Суммарная ошибка равна удвоенной оценке EER, однако и она не характеризует свойства системы распознавания, поскольку сдвиг порога принятия решений в сторону уменьшения

вероятности пропуска самозванца приводит в экспоненциальному росту вероятности отказа целевому диктору.

Достигнутые характеристики систем идентификации диктора могут удовлетворять требованиям практической применимости в условиях малой вероятности вторжения самозванца, малой стоимости ошибки, или в случаях, когда окончательное решение принимается экспертом. При управлении финансовыми операциями или доступе к конфиденциальной информации стоимость ошибки велика, и необходимо значительно уменьшить вероятность пропуска самозванца при сохранении вероятности отказа целевому диктору в психологически приемлемых пределах.

Литература

1. Adami A. (2007). Modeling prosodic differences for speaker recognition. Speech Communication, v. 49, N4, 277–291.
2. Adami A., Mihaescu R., Reynolds D., Godfrey J. (2003). Modeling prosodic dynamics for speaker recognition. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 788–791.
3. Airas M., Alku P. (2006). Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalized amplitude quotient. Phonetica 63 (1), 26–46.
4. Altincay H., Demirekler M. (2003). Speaker identification by combining multiple classifiers using Dempster–Shafer theory of evidence. Speech Communication, v.41, N4, 531–547.
5. Andrews W., Kohler M., Campbell J. (2001). Phonetic speaker recognition. In: Proc. Seventh European Conf. on Speech Communication and Technology (Eurospeech 2001), Aalborg, Denmark, September 2001, 2517–2520.
6. Anguera X., Bonastre J.-F. (2010). A Novel Speaker Binary Key Derived from Anchor Models. Interspeech, 2118-2121.
7. Ashour G., Gath I. (1999). Characterization of speech during imitation. In: Proc. Sixth European Conf. on Speech Communication and Technology (Eurospeech 1999), Budapest, Hungary, September 1999, 1187–1190.
8. Atal B. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Amer., v.55, N6, 1304–1312.
9. Atal B. (1976). Automatic recognition of speakers by their voice. Proc. IEEE, V.64, N4, 460-475.
10. Auckenthaler R., Mason J. (2001). Gaussian selection applied to textindependent speaker verification. In: Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey), 83–88.
11. Barras C., Xuan Z., Meignier S., Gauvian J. (2006). Multistage speaker diarization of broadcast news. IEEE Trans. Audio, Speech and Language Processing, v.14, 1505–1512.
12. Bartkova K., Gac D.L., Charlet D., Jouvet D. (2002). Prosodic parameter for speaker identification. Proc. Internat. Conf. on Spoken Language Processing (ICSLP), 1197–1200.
13. BenZeghiba M., Bourlard H. (2003). On the combination of speech and speaker recognition. In: Proc. Eighth European Conf. on Speech Communication and Technology (Eurospeech), 1361–1364.
14. Besacier L., Bonastre J.-F. (2000). Subband architecture for automatic speaker recognition. Signal Process., v.80, 1245–1259.
15. Besacier L., Bonastre J., Fredouille C. (2000). Localization and selection of speaker-specific information with statistical modeling. Speech Communication, v.31, 89–106.
16. Bimbot F., Blomberg M., Boves L., Genoud D., Hutter H.-P., Jaboulet C., Koolwaaij J., Lindberg J., Pierrot J.-B. (2000). An overview of the CAVE project research activities in speaker verification. Speech Communication, v. 31, 155-180.
17. Bimbot F., Magrin-Chagnolleau I., Mathan L. (1995). Second-order statistical measures for text-independent speaker identification. Speech Communication, v.17, 177–192.
18. Bimbot F., Bonastre J.-F., Fredouille C., Gravier G., Magrin-Chagnolleau I., Meignier S., Merlin T., Ortega-Garcia J., Petrovska-Delacretaz D., Reynolds D. (2004). A tutorial on text-independent speaker verification. EURASIP J. Appl. Signal Process., v.4, 430–451.
19. Bishop C. (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, LLC, New York.
20. Bocklet T., Shriberg E. (2009). Speaker recognition using syllable-base constraints for cepstral frame selection. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 4525–4528.
21. Boë L.J. (2000). Forensic voice identification in France. Speech Communication, v. 31, 205-224.
22. Bonastre J.-F., Matrouf D., Fredouille C. (2007). Artificial impostor voice transformation effects on false acceptance rates. In: Proc. Interspeech 2007 (ICSLP), Antwerp, Belgium, August 2007, 2053–2056.

23. Braun A. (1996). Age estimation by different listener groups. The J. of Speech Language and the Law, v.3, N1, 65-73.

24. Brugnara F., Falavigna D., Omologo M. (1993). Automatic segmentation and labeling of speech based on hidden markov models, Speech Communication, v.12, N4, 357–370,.

25. Burget L., Matějka P., Schwarz P., Glembek O., Černockŷ J. (2007). Analysis of feature extraction and channel compensation in a GMM speaker recognition system. IEEE Trans. Audio, Speech Language Process., v.15, N7, 1979–1986.

26. Burton D. (1987). Text-dependent speaker verification using vector quantization source coding. IEEE Trans. Acoustics, Speech, Signal Process., v.35, N2, 133–143.

27. Byrne C., Foulkes P. (2004). The "mobile phone effect" on vowel formants. Speech,Language and Law, v.11, N1, 83-102.

28. Campbell J. (1997). Speaker recognition: a tutorial. Proc. IEEE, v.85, N9, 1437–1462.

29. Campbell W., Assaleh K., Broun C. (2002). Speaker recognition with polynomial classifiers. IEEE Trans. Speech Audio Process., v.10, N4, 205–212.

30. Campbell W., Campbell J., Reynolds D., Jones D., Leek T. (2004). Phonetic speaker recognition with support vector machines. In: Thrun, S., Saul, L., Schokopf, B. (Eds.), . In: Advances in Neural Information Processing Systems, v.16. MIT Press, Cambridge, MA.

31. Campbell J.P., Reynolds D.A., Dunn R.B. (2003). Fusing high- and lowlevel features for speaker recognition. In: Proc. Eurospeech, 2665–2668.

32. Campbell W., Campbell J., Reynolds D., Singer E., Torres-Carrasquillo P. (2006a). Support vector machines for speaker and language recognition. Comput. Speech Lang., v.20, N2–3, 210–229.

33. Campbell W. M., Karam (2010). Simple and Efficient Speaker Comparison using Approximate KL Divergence. Interspeech, 362-365.

34. Campbell J. P., Shen W., Campbell W.M., Schwartz R., Bonastre J.-F., Matrouf D. (2009). Forensic Speaker Recognition: A need for caution. IEEE Signal Processing Magazine, v.95.

35. Campbell W., Sturim D., Reynolds D. (2006b). Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process. Lett., v.13, N5, 308–311.

36. Carey M., Parris E., Lloyd-Thomas H., Bennett S. (1996). Robust prosodic features for speaker identification. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP), 1800–1803.

37. Castaldo F., Colibro D., Dalmasso E., Laface P., Vair C. (2007). Compensation of nuisance factors for speaker and language recognition. IEEE Trans. Audio, Speech Language Process., v.15, N7, 1969–1978.

38. Cerrato L., Falcone M., Paoloni A. (2000). Subjective age estimation of telephonic voices, Speech Communication, v. 31, 107-112.

39. Champod Ch., Meuwly D. (2000). The inference of identity in forensic speaker recognition. Speech Communication, v. 31, 193-2003.

40. Chappell D., Hansen, J. (1998). Speaker-specific pitch contour modeling and modification. Proc. ICASSP, v. 1, 885–888.

41. Charbuillet C., Gas B., Chetouani M., Zarader J. (2006). Filter bank design for speaker diarization based on genetic algorithms. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP2006), Vol. 1, Toulouse, France, May 2006, 673–676.

42. Charlet D., Jouvet D., Collin J. (2000). An alternative normalization scheme in HMM-based text-dependent speaker verification, Speech Communication, v. 31 113-120.

43. Chaudhari U., Navratil J., Maes S. (2003). Multigrained modeling with pattern specific maximum likelihood transformations for text-independent speaker recognition. IEEE Trans. Speech Audio Process., v.11, N1, 61–69.

44. Chen G., Feng X., Shue Y.-L., Alwan A. (2010). On Using Voice Source Measures in Automatic Gender Classification of Children's Speech. Interspeech, 673-676.

45. Chen K., Wang L., Chi H. (1997). Methods of combining multiple classifiers with different features and their applications to text-independent speaker recognition. Internat. J. Pattern Recognition Artif. Intell., v.11, N3, 417–445.

46. Cheng S. S., Wang H. M., Fu C. (2010). BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization. IEEE Trans. on Audio, Speech and Language Processing, V.18, N1,141-157.

47. Chibelushi C.C., Deravi F., Mason J.S.D. (2002). A Review of speechbased bimodal recognition. IEEE Trans. on Multimedia, v.4.

48. Dahan D., Bernard J.M. (1996). Interspeaker variability in emphatic accent production in French. Lang. Speech, v. 39, N4, 341–374.

49. Dang J., Honda K. (1996a). An improved vocal tract model of vowel production implementing piriform fossa resonance and transvelar nasal coupling. In: Proc. ICSLP1996, 965–968.
50. Dang J., Honda K. (1996b). Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation. J. Acoust. Soc. Am., v.100, 3374–3383.
51. Dang J., Honda K. (1997). Acoustic characteristics of the piriform fossa in models and humans. J. Acoust. Soc. Am., v.101, 456–465.
52. Davis S., Mermelstein P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech, Signal Process., v.28, N4, 357–366.
53. Dehak N., Chollet G. (2006). Support vector GMMs for speaker verification. In: Proc. IEEE Odyssey: the Speaker and Language Recognition Workshop (Odyssey).
54. Deller J., Hansen J., Proakis J. (2000). Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York.
55. Dempster A., Larid N., Rubin D. (1979). Maximum likelihood estimation from incomplete data. Journal of the Royal Statistical Society B 39 (1), 1-38.
56. Doddington G.R. (1985). Speaker recognition – Identifying people by their voices. Proc. of IEEE, v.73, N11, 1651-1664.
57. Doddington G. (2001). Speaker recognition based on idiolectal differences between speakers. In: Proc. Seventh European Conf. on Speech Communication and Technology (Eurospeech), 2521–2524.
58. Doddington G.R., Przybocky M.A., Martin A.F., Reynolds D.A. (2000). The NIST speaker recognition evaluation – Overview, methodology, sytrems, results, perspective. Speech Communication, v. 31, 225-254.
59. Drugman T., Dutoit T. (2010). On the Potential of Glottal Signatures for Speaker Recognition. Interspeech, 2106-2109.
60. Drygajlo A. (2007). Forensic Automatic Speaker Recognition. IEEE Signal Processing Magazine, v.32.
61. Faltlhauser R., Ruske G. (2001). Improving speaker recognition performance using phonetically structured gaussian mixture models. In: Proc. Seventh European Conf. on Speech Communication and Technology (Eurospeech), 751–754.
62. Farrell K., Mammone R., Assaleh K. (1994). Speaker recognition using neural networks and conventional classifiers. IEEE Trans. Speech Audio Process., v.2, N1, 194–205.
63. Faundez-Zanuy M. (2004). On the vulnerability of biometric security systems. IEEE Aerospace Electron. Syst. Mag., v.19, N6, 3–8.
64. Faundez-Zanuy M. (2005). Data fusion in biometrics. IEEE Aerospace Electron. Syst. Mag., v.2,0 N1, 34–38.
65. Faundez-Zanuy M., Hagmüller M., Kubin G. (2006). Speaker verification security improvement by means of speech watermarking. Speech Communication, v.48, 1608–1619.
66. Federico M. (1996). Bayesian estimation methods for n-gram language model adaptation. In: ICSLP, pp. 279–282.
67. Fergani B., Davy M., Houacine A. (2008). Speaker diarization using one-class support vector machines. Speech Communication, v.50, 355–365.
68. Ferrer L., Shriberg E., Kajarekar S., Sönmez K. (2007). Parameterization of prosodic feature distributions for SVM modeling in speaker recognition. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2007), v.4, 233–236.
69. Feustel T.C., Velius G.A., Logan R.J. (1989). Human and machine performance on speaker identity verification. Speech Tech'89, 169-170.
70. Forsyth M. (1995). Discriminating observation probability (DOP) HMM for speaker verification. Speech Communication, v.17, N1-2, 117-129.
71. Freund Y., Schapire R.E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting . Journal of Computer and System Sciences, v. 55. 119-139.
72. Friedland A. G., Vinyals B. O., Huang C. Y. (2009). Fusing short term and long term features for improved speaker diarization. ICASSP, 4077-4080.
73. Furui S., Itakura F., Saito S. (1972). Talker recognition by longtime averaged speech spectrum. Trans. IECE 55-A, v.1, 549-556.
74. Furui S. (1981). Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoustics, Speech Signal Process., v.29, N2, 254–272.

75. Gauvain J.L., Lee C.H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Trans. Speech Audio Process., v.2, 291–298.

76. Ghosh P. K., Narayanan Sh. S. (2011). Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter. Speech Communication, v. 53, 98–109.

77. Gudnason J., Brookes M. (2008). Voice source cepstrum coefficients for speaker identification. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 4821–4824.

78. Hansen E., Slyh R., Anderson T. (2004). Speaker recognition using phoneme-specific GMMs. In: Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey), 179–184.

79. Harmegnies B., Landercy A. (1988). Intra-speaker variability of the long term spectrum. Speech Communication, v.7, 81-86.

80. Hautamäki V., Kinnunen T., Nosratighods M., Lee K.-A., Ma B., Li H. (2010). Approaching Human Listener Accuracy with Modern Speaker Verification. Interspeech, 1473-1476.

81. He J., Liu L., Palm G. (1999). A discriminative training algorithm for VQ-based speaker identification. IEEE Trans. Speech Audio Process., v.7, N3, 353–356.

82. Hébert M., Heck L. (2003). Phonetic class-based speaker verification. In: Proc. Eighth European Conf. on Speech Communication and Technology (Eurospeech), 1665–1668.

83. Heck L.P., Konig Y., Sönmez M.K., Weintrab M. (2000). Robustness to telephone handset distortion in speaker recognition by discriminative feature design. Speech Communication, v. 31, 181-192.

84. Hennbert J., Melin H., Petrovska D., Genoud D. (2000). POLYCOST: A telephone-speech database for speaker recognition. Speech Communication, v. 31, 265-270.

85. Hermansky H., Morgan N. (1994). RASTA processing of speech. IEEE Trans. Speech Audio Process., v.2, N4, 578–589.

86. Heuvel H. van den, Cranen B., Rietveld T. (1996). Speaker variability in the coarticulation of /a,i,u/, Speech Communication, v.18, 113- 130.

87. Ho T.K., Hull J., Srihari S. (1994). Decision combination in multiple classifiers systems. IEEE Trans. Pattern Anal. Machine Intell., v.16, 66-75.

88. Hollien H. (2002). Forensic voice identification. Academic Press.

89. Huang X., Acero A., Hon H.-W. (2001). Spoken Language Processing: a Guide to Theory, Algorithm, and System Development. Prentice-Hall, New Jersey.

90. Itoh K. (1992). Perceptual analysis of speaker identity. In: Saito, S. (Ed.), Speech Science and Technology. IOS Press, 133-145.

91. Jacewicz E., Fox R. A. (2010). Between-speaker and within-speaker variation in speech tempo of American English. J. Acoust. Soc. Am. 128, N 2, 839–850.

92. Jin Q., Schultz T., Waibel A. (2002). Speaker identification using multilingual phone strings. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002), v. 1, 145–148.

93. Johnson N.F., Jajodia S. (1998). Exploring Steganography: seeing the unseen. IEEE Computer (February), 26–34.

94. Kajarekar S., Ferrer L., Shriberg E., Sönmez K., Stolcke A., Venkataraman A., Zheng J. (2005). Speech recognition performance comparison between DSR and AMR transcoded speech. In: Proc. ICASSP, v. 1, pp. 173– 176.

95. Karam Z., Campbell W. (2007). A new kernel for SVM MLLR based speaker recognition. In: Proc. Interspeech 2007 (ICSLP), Antwerp, Belgium, August 2007, 290–293.

96. Karpov E., Kinnunen T., Fränti, P. (2004). Symmetric distortion measure for speaker recognition. In: Proc. Ninth Internat. Conf. on Speech and Computer (SPECOM), 366–370.

97. Kenny P. (2006). Joint factor analysis of speaker and session variability: theory and algorithms. Technical Report CRIM-06/08-14.

98. Kenny P., Boulianne, G., Dumouchel P. (2005). Eigenvoice modeling with sparse training data. IEEE Trans. Speech Audio Process., v.13, 345–354.

99. Kenny P., Boulianne G., Ouellet P., Dumouchel P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. Audio Speech Lang. Process., v.15, 1435–1447.

100. Kenny P., Boulianne G., Ouellet P., Dumouchel P. (2007). Speaker and session variability in GMM-based speaker verification. IEEE Trans. Audio, Speech Language Process., v.15, N4, 1448–1460.

101. Kenny P., Ouellet P., Dehak N., Gupta V., Dumouchel P. (2008). A study of inter-speaker variability in speaker verification. IEEE Trans. Audio, Speech Language, v.16, N5, 980–988.

102. Kersta L.G. (1962).Voiceprint Identification, Nature, v.196, 1253-1257.

103. Kinnunen T., Gonzälez-Hautamäki, R. (2005). Long-term f0 modeling for text-independent speaker recognition. In: Proc. 10th Internat. Conf. on Speech and Computer (SPECOM'2005), 567–570.

104. Kinnunen T., Karpov E., Fränti P. (2006b). Real-time speaker identification and verification. IEEE Trans. Audio, Speech Language Process., v.14, N1, 277–288.

105. Kinnunen T., Li H. (2010). An overview of text-independent speaker recognition: from features to supervectors," Speech Communication, v.52, N1, 12–40.

106. Kinnunen T., Saastamoinen J., Hautamäki V., Vinni M., Fränti P. (2009). Comparative evaluation of maximum a posteriori vector quantization and Gaussian mixture models in speaker verification. Pattern Recognition Lett., v.30, N4, 341–347.

107. Kimbal O., Schmidt M., Gish H., Waterman J. (1997). Speaker verification with limited enrollment data. In: Eurospeech'97, 967–970.

108. Kitamura T. (2008). Acoustic analysis of imitated voice produced by a professional impersonator. Proc. Interspeech 2008, 813–816.

109. Kittler J., Hatef M., Duin R., Matas J. (1998). On combining classifiers. IEEE Trans. Pattern Anal. Machine Intell., v.20, N3, 226–239.

110. Klusacek D., Navratil J., Reynolds D.A., Campbell J.P. (2003). Conditional pronunciation modeling in speaker detection. In: Proc. ICASSP, v. 4, pp. 804–807.

111. Kolano G., Regel-Brietzmann P. (1999). Combination of vector quantization and Gaussian mixture models for speaker verification. In: Proc. Sixth European Conf. on Speech Communication and Technology (Eurospeech), 1203–1206.

112. Kuhn R., Junqua J.C., Nguyen P. Niedzielski N. (2000). Rapid speaker adaptation in eigenvoice space. IEEE Trans. Speech Audio Process., v.8, 695–707.

113. Künzel H.I. (2001). Beware of the "telephone effect": the influence of telephone transmission on the measurement of formant frequencies. Forensic Linguistics, v.8, N1, 80-99.

114. Kuo J.-W., Wang H.-M. (2006). Minimum Boundary Error Training for Automatic Phonetic Segmentation. Interspeech, 1217-1220.

115. Kuo J.-W., Lo H.-Y., Wang H.-M. (2007). Improved HMM/SVM methods for automatic phoneme segmentation. Interspeech, 2057–2060.

116. Kuwabara H., Sagisaka Y. (1995). Acoustic characteristics of speaker individuality: Control and Conversion. Speech Communication, v.16, 165-173.

117. Lamel L.F., Gauvain J.L. (2000). Speaker verification over the telephone. Speech Communication, v. 31, 141-154.

118. Laskowski K., Jin Q. (2009). Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2009), Taipei, Taiwan, April 2009, 4541–4544.

119. Lass N.J., Philips J.K., Bruchey C.A. (1980). The effect of filtered speech on speaker height and weight identification. J. of Phonetics, N8, 91-100.

120. Lavner Y., Gath I., Rosenhouse J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. Speech Communication v.30, 9-26.

121. Lee K., You C., Li H., Kinnunen T., Zhu D. (2008). Characterizing speech utterances for speaker verification with sequence kernel SVM. In: Proc. Ninth Interspeech (Interspeech 2008), Brisbane, Australia, September 2008, 1397–1400.

122. Leggetter C., Woodland P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. Comput. Speech Lang., v.9, 171–185.

123. Lei Y., Hansen J. H.L. (2011). Mismatch modeling and compensation for robust speaker verification. Speech Communication, v.53, 257–268

124. Leng Y. R., Tran H. D., Kitaoka N., Li H. (2010). Selective Gammatone Filterbank Feature for Robust Sound Event Recognition. Interspeech, 2246-2249.

125. Leung K., Mak M., Siu M., Kung S. (2006). Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification. Speech Communication, v.48, N1, 71–84.

126. Linde Y., Buzo A., Gray R. (1980). An algorithm for vector quantizer design. IEEE Trans. Communication, v.28, N1, 84–95.

127. Littlestone N., Warmuth M. (1994). The weighted majority algorithm. Information and Computation., v. 108, 212-261.

128. Lu X., Dang J. (2007). An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Communication, v.50, N4, 312–322.

129. McLachlan G., Peel D. (2000). Finite Mixture Models. New York: John Wiley & Sons Inc.

130. Mak B., Ho S., Hsiao R., Kwok J.T. (2005). Kernel eigenvoice speaker adaptation. IEEE Trans. Speech Audio Process., v.13, 984–992.

131. Mak M.-W., Hsiao R., Mak B. (2006). A comparison of various adaptation methods for speaker verification with limited enrollment data. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), v. 1, 929–932.

132. Mak M.-W., Rao W. (2011). Utterance partitioning with acoustic vector resampling for GMM–SVM speaker verification. Speech Communication, v.53, 119–130.

133. Malayath N., Hermansky H., Kajarekar S., Yegnanarayana B. (2000). Data-driven temporal filters and alternatives to GMM in speaker verification. Digital Signal Process., v.10, N1–3, 55–74.

134. Mariénthoz J., Bengio S. (2002). A comparative study of adaptation methods for speaker verification. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP), 581–584.

135. Markel J., Oshika B., Gray Jr. A.H. (1977). Long-term feature averaging for speaker recognition. IEEE Trans. Acoustics, Speech, Signal Process., v.25, N4, 330–337.

136. Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M. (1997). The DET curve in assessment of detection task performance. In: Proc. Fifth European Conf. on Speech Communication and Technology (Eurospeech), 1895–1898.

137. Matsui T., Furui S. (1993). Concatenated phoneme models for textvariable speaker recognition. In: Proc. ICASSP, v. 1, pp. 391–394.

138. Matsumoto H., Hiki S., Sone T., Nimura T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates, IEEE Trans. AU, v. AU- 21, 428-436.

139. Meignier S., Moraru D., Fredouille C., Bonastre J.-F., Besacier L. (2006). Step-by-step and integrated approaches in broadcast news speaker diarization. Computer Speech and Language, v.20, N 2-3, 303–330.

140. Miyajima C., Watanabe H., Kitamura T., Katagiri S. (1999). Discriminative feature extraction – Optimization of Mel-cepstral features using second-order all-pass warping function. Proc. EUROSPEECH, II-779–I-782.

141. Müller C. (Ed.) (2007). Speaker Classification I: Fundamentals, Features, and Methods. Lecture Notes in Computer Science, v. 4343. Springer.

142. Murthy H.A., Beaufays F., Heck K.P., Weintraub M. (1999). Robust text-independent speaker identification over telephone channels. IEEE Trans. Speech and Audio Process., v.7, 554-558.

143. Murty K., Yegnanarayana B. (2006). Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Process. Lett., v.13, N1, 52–55.

144. Naik J., Netsch L., Doddington G. (1989). Speaker verification over long distance telephone lines. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 524–527.

145. Naonori U., Ryohei N., Ghahramani Z., Hinton G.E. (2000). SMEM Algorithm for Mixture Models. Neural Computation, v.12, N9, 2109-2128.

146. Navratil J., Jin Q., Andrews W., Campbell J. (2003). Phonetic speaker recognition using maximum likelihood binary decision tree models. In: Proc. ICASSP, v. 4, pp. 796–799.

147. Neiman G.S., Applegate J.A. (1990). Accuracy if listener judgemnets of perceived age relative to chronological age in adults. Folia Phomiatrica, v.42, 327-340.

148. Nolan F. (1983). The Phonetic Bases of Speaker Recognition. Cambridge University Press, Cambridge.

149. Nolan F., Oh T. (1996). Identical twins, different voices. Forensic Linguistics, v.3, N1, 39–49.

150. Ortega-Garcia J., Gonzalez-Rodriguez J., Marreo-Aguiar V. (2000). AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. Speech Communication, v. 31, 225-264.

151. Pardo J. M., Anguera X., Wooters C. (2007). Speaker diarization for multiple-distant-microphone meetings using several sources of information. IEEE Trans. on Computers, v.56, N9, 1212- 1224.

152. Park A., Hazen T. (2002). ASR dependent techniques for speaker identification. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP), 1337–1340.

153. Patterson R. D., Holdsworth J. (1996). A functional model of neural activity patterns and auditory images. Advances in Speech, Hearing and Language Processing, v. 3, 547–563.

154. Pelecanos J., Sridharan S. (2001). Feature warping for robust speaker verification. In: Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey), Crete, Greece, June 2001, 213–218.

155. Pelecanos J., Myers S., Sridharan S., Chandran V. (2000). Vector quantization based Gaussian modeling for speaker verification. In: Proc. Internat. Conf. on Pattern Recognition (ICPR), 3298–3301.

156. Pellom B., Hansen J. (1999). An efficient scoring algorithm for Gaussian mixture model based speaker identification. IEEE Signal Process. Lett., v.5, N11, 281–284.

157. Peskin B., Navratil J., Abramson J., Jones D., Klusäček D., Reynolds D., Xiang B. (2003). Using prosodic and conversational features for high-performance speaker recognition: report from JHU WS'02. Proc. ICASSP, v. 4, 792–795.

158. Plumpe M., Quatieri T., Reynolds D. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. Speech Audio Process., v.7, N5, 569–586.

159. Prasanna S. R. M., Govind D. (2010). Analysis of Excitation Source Information in Emotional Speech. Interspeech, 781-784 Рамишвили Г.С. (1981). Автоматическое опознавание говорящего по голосу. М: Радио и Связь, 222 с.

160. Reddy M S. H., Prahallad K., Gangashetty S.V., Yegnanarayana B. (2010). Significance of Pitch Synchronous Analysis for Speaker Recognition using AANN Models. Interspeech, 669-672.

161. Reynolds D. (1995). Speaker identification and verification using Gaussian mixture speaker models. Speech Communication, v.17, 91–108.

162. Reynolds D., Rose R. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process., v.3, 72–83.

163. Reynolds D.A. (1997a). Comparison of background normalization methods for text-independent speaker verification. In: Proc. Eurospeech'97, 963–966.

164. Reynolds D.A. (1997b). HTIMIT and LLHDB: speech corpora for the study of handset transducer effects. In: Proc. ICASSP, v. 2, 1535–1538.

165. Reynolds D., Quatieri T., Dunn R. (2000). Speaker verification using adapted gaussian mixture models. Digital Signal Process., v.10, N1, 19–41.

166. Reynolds D., Andrews W., Campbell J., Navratil J., Peskin B., Adami A., Jin Q., Klusacek D., Abramson J., Mihaescu R., Godfrey J. Jones D., Xiang B. (2003). The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 784–787.

167. Reynolds D., Kenny P., Castaldo F. (2009). A study of new approaches to speaker diarization. Interspeech, 6–10.

168. Ромашкин Ю.Н., Петров Ю.О. (2009). Распознавание пола диктора на основе GMM-модели голоса. Речевые технологии, №2, 31-38.

169. Rose P. (2002). Forensic Speaker Identification. Taylor & Francis, London.

170. Rosenberg A. (1976). Automatic speaker recognition. Proc. IEEE, v. 64, N4, 475-478.

171. Rosenberg A., Siohan O., Parthasarathy S. (2000). Small group speaker identification with common password phrases. Speech Communication, v. 31, 131-140.

172. Schmidt-Nielsen A., Crystal Th. H. (2000). Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data. Digital Signal Processing, v. 10, 249–266.

173. Shriberg E., Ferrer L., Kajarekar S., Venkataraman A. Stolcke A. (2005). Modeling prosodic feature sequences for speaker recognition. Speech Communication,v.46, N3–4, 455–472.

174. Shuwmaker M.B., Hapner E.R., Gilman M., Klein A.M., Johns M.M. (2010). Analysis of voice change during cellular phone use: a blinded controlled study. J. of Voice, v.23, N3, 308-313.

175. Siohan O., Chesta, C., Lee, C.H. (2001). Joint maximum a posteriori adaptation of transformation and HMM parameters. IEEE Trans. Speech Audio Process. 9 (4), 417–428.

176. Sönmez M., Heck L., Weintraub M., Shriberg E. (1997). A lognormal tied mixture model of pitch for prosody-based speaker recognition. In: Proc. Fifth European Conf. on Speech Communication and Technology (Eurospeech), 1391–1394.

177. Sönmez K., Shriberg E., Heck L., Weintraub M. (1998). Modeling dynamic prosodic variation for speaker verification. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP 1998), 3189–3192.

178. Soong F., Rosenberg A. (1988). On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans. Acoustics, Speech Signal Process., v.36, N6, 871–879.

179. Soong F.K., Rosenberg A.E., Juang B.-H., Rabiner L.R. (1987). A vector quantization approach to speaker recognition. AT & T Technical J., v.66, 14–26.

180. Сорокин В.Н. (1992). Синтез речи. Наука, Москва.

181. Sorokin V.N., Makarov I.S. (2008). Gender recognition from vocal source. Acoustical Physics, v. 54, N4, 571-578.

182. SorokinV.N., Tsyplikhin A.I. (2010). Speaker verification using the spectral and time parameters of voice signal. Journal of Communications Technology and Electronics, v.55, N12, 1561-1574.

183. Stolcke A., Kajarekar S., Ferrer L., Shriberg E. (2007). Speaker recognition with session variability normalization based on MLLR adaptation transforms. IEEE Trans. Audio, Speech Language Process., v.15, N7, 1987–1998.

184. Su L.-S., Li K.P., Fu K.S. (1974). Identification of speakers by use of nasal coarticulation, J. Acoust. Sot. Amer., v. 56, 1867-1882.

185. Sussman J., Dalston E., Gumbert S. (1998). The effect of speaking style on a locus equation characterization of stop place articulation. Phonetica, v.55, N4, 204–255.

186. Suzuki H., Nakai T., Dang J., Lu C. (1990). Speech production model involving subglottal structure and oral–nasal coupling through closed velum. In: Proc. ICSLP90, 437–440.

187. Takemoto H., Adachi S., Kitamura T., Mokhtari P., Honda K. (2006). Acoustic roles of the laryngeal cavity in vocal tract resonance. J. Acoust. Soc. Am., v.120, 2228–2239.

188. Thévenaz P., Hügli H. (1995). Usefulness of the LPC-residue in textindependent speaker verification. Speech Communication, v.17, N1–2, 145–157.

189. Thyes O., Kuhn R., Nguyen P., Junqua J.C. (2000). Speaker identification and verification using eigenvoices. In: Proc. ICSLP, v. 2, pp. 242–245.

190. Tranter S., Reynolds D.A. (2006). An overview of automatic speaker diarization systems. IEEE Trans. Audio, Speech Language Process., v.14, N5, 1557–1565.

191. Tsao Y.-C., Weismer G. (1997). Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component, J. Speech Lang. Hear. Res., v.40, 858–866.

192. TWGFAST (1997). Technical working group on friction ridge analysis, study and technology (TWGFAST) proposed guidelines. J. of Forensic Identification, v.47, N4, 423-437.

193. Van Lancker, D., Kreiman, J., and Emmorey, K. (1985). Familiar voice recognition: patterns and parameters. Part I. Recognition of backwards voices. J. Phonetics v.13, 19–38.

194. van Lancker D., Kreiman J., Emmory R. (1985a). Familiar voice recognition: pattern and parameters. Part 2: Recognition of rate-altered voices. J. of Phonetics, v.13, 39-52).

195. Vapnik V.N. (1998). Statistical Learning Theory. New York: Wiley.

196. Vogt R., Sridharan S. (2008). Explicit modeling of session variability for speaker verification. Comput. Speech Lang., v.22, N1, 17–38.

197. Wan V., Renals S. (2005). Speaker verification using sequence discriminant support vector machines. IEEE Trans. Speech Audio Process., v.13, N2, 203–210.

198. Wang L., Kitaoka N., Nakagawa S. (2007). Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM. Speech Communication, v. 49, 501–513.

199. Wang D., Vogt R., Sridharan S. (2010). Bayes Factor Based Speaker Segmentation for Speaker Diarization. Interspeech,1405-1408.

200. Weber F., Manganaro L., Peskin B., Shriberg E. (2002). Using prosodic and lexical information for speaker identification. In: Proc. ICASSP, v. 1, 141–144.

201. Wester M. (2010). Cross-lingual talker discrimination. Interspeech, 1253-1256.

202. Wu T., Duchateau J., Martens J.-P., Compernolle D. V. (2010). Feature subset selection for improved native accent identification. Speech Communication, v.52, 83–98.

203. Xiang B., Chaudhari U., Navratil J., Ramaswamy G., Gopinath R. (2002). Short-time Gaussianization for robust speaker verification. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), v. 1, 681–684.

204. Xiong Z., Zheng T., Song Z., Soong F., Wu W. (2006). A tree-based kernel selection approach to efficient Gaussian mixture modeluniversal background model based speaker identification. Speech Communication, v.48, 1273–1282.

205. Yegnanarayana B., Kishore S. (2002). AANN: an alternative to GMM for pattern recognition. Neural Networks, v.15, 459–469.

206. Yegnanarayana B. Satyanarayana P. (2000). Enhancement of reverberant speech using LP residual signal. IEEE Trans. Acoust., Speech, Signal Processing, v. 8, N. 3, 267–281.

207. Yoma N. B., Garreton C., Molina C., Huenupan F. (2008). Unsupervised intra-speaker variability compensation based on Gestalt and model adaptation in speaker verification with telephone speech. Speech Communication, v.50, 953–964.

208. You C., Lee K., Li H. (2009). An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition. IEEE Signal Process. Lett., v.16, N1, 49–52.

209. Yuo K.-H., Wang H.-C. (1999). Joint estimation of feature transformation parameters and Gaussian mixture model for speaker identification. Speech Communication, v.28, N3, 227–241.

210. Zhang W.-Q., Deng Y., He L., Liu J.(2010). Variant Time-Frequency Cepstral Features for Speaker Recognition. Interspeech, 2122-2125.

211. Zhang Sh.-X, Mak M.-W. (2009). A new adaptation approach to high-level speaker-model creation in speaker verification. Speech Communication, v. 51, 534-550.

212. Zilca R. (2002). Text-independent speaker verification using utterance level scoring and covariance modeling. IEEE Trans. Speech Audio Process., v.10, N6, 363–370.

213. Рамишвили Г.С. (1981). Автоматическое опознавание говорящего по голосу. М: Радио и Связь, 222 с.