Математические структуры и моделирование
2011, вып. 22, с. 48–54

УДК 534.4

ПОДХОДЫ К ЗАДАЧЕ ИДЕНТИФИКАЦИИ ДИКТОРА

О. А. Вишнякова, Д. Н. Лавров

Представлен обзор систем идентификации по голосу, алгоритмов распознавания дикторов. Проводится анализ информативных признаков, способов построения эталонов и алгоритмов принятия решения.

Введение

Интерес к системам идентификации обусловлен широким кругом практических приложений: проверка прав доступа к различным системам (базам данных, каналам связи, помещениям, устройствам и механизмам, банковским счетам и т. д.), криминалистическая экспертиза.

Преимущества именно голосовой идентификации следуют из характеристик голоса: не отчуждаем от человека; не требует непосредственного контакта; не требует сложных технических устройств. Голос диктора, а как следствие и сам речевой сигнал уникален ввиду специфики физиологического строения его артикуляторного аппарата и специфики его речи. Это обусловливает интерес к нему как биометрическому объекту. Только в некоторых весьма редких случаях указанная уникальность может не иметь места (например, однояйцовые близнецы, воспитанные в одинаковых условиях).

1. Классификация задач определения диктора

Классически задача идентификации выглядит следующим образом: имеется ограниченная и строго контролируемая группа пользователей системы. Системы идентификации диктора выносят решение о том, что диктор принадлежит к выбранной группе (или подгруппе) дикторов, модели которых хранятся в её базе моделей, и указывают конкретного диктора (рис. 1). При этом качество системы характеризуется средней вероятностью правильной идентификации. При такой формулировке задачи исключена ситуация возможного злоумышленника, однако среди возможных применений систем распознавания дикторов ситуации с замкнутыми группами возникают редко [2].

Омский государственный университет им. Ф. М. Достоевского.
E-mail: olga@infotekorg.ru

Рис. 1. Общая схема идентификации диктора

На практике чаще встречается задача верификации, для решения которой строится система, которая выносит решение о том, что диктор, голосовой сигнал которого предъявлен системе в качестве образца, соответствует его модели голоса, хранимой в базе голосовых моделей зарегистрированных пользователей и обозначенной уникальным PIN-кодом, или не соответствует. Диктор в первом случае может быть обозначен как «свой», а во втором — как «чужой» (рис. 2). Качество принимаемого системой решения характеризуется ошибками 1-го и 2-го рода (FRR и FAR) [5].

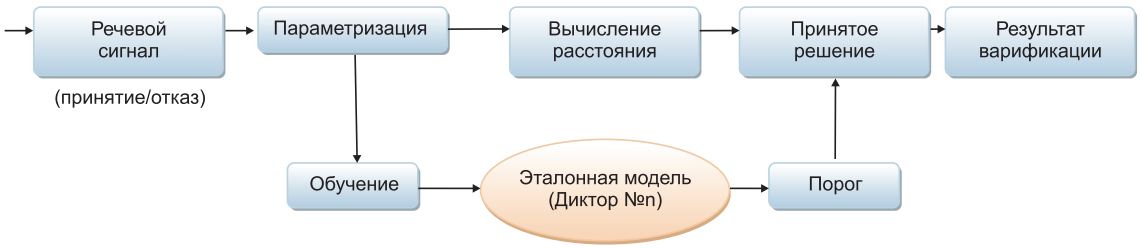

Рис. 2. Общая схема верификации диктора

В общем же случае задача распознавания диктора сводится к открытой идентификации, при которой пользователь не объявляет свою индивидуальность. Система должна сверить поступивший речевой сигнал со всеми речевыми эталонами зарегистрированных пользователей. Таким образом, задача открытой идентификации совпадает с задачей многократной верификации.

В настоящее время используются как текстозависимые, так и текстонезависимые системы распознавания дикторов. Текстозависимая система работает по парольным фразам (статический режим). Обобщённым случаем для этого режима является тексто-подсказанный режим (динамический), когда система в случайном порядке предлагает диктору сказать фразу из заданного набора, причём диктор на этапе обучения ввёл соответствующие фразы в систему. Содержание фраз может выбираться пользователем или системой. Это свойство системы позволяет пользователю периодически менять свой голосовой пароль, обеспечивая ещё большую надёжность верификации. В отличие от предыдущей, текстонезависимая система работает с использованием произвольной речи. Диктору не нужно помнить какую-то определённую парольную фразу. Очевидно, что совпадение лингвистической формы двух сравниваемых речевых сообщений облегчает процесс идентификации. В процессе решения задачи идентификации каждому диктору ставится в соответствие некоторый эталон — набор уникальных признаков — для дальнейшей классификации и принятия решения алгоритмами распознавания.

Таким образом, задача разбивается на три относительно независимые части:

1. Выделение информативных признаков (параметризация речевого сигнала);
2. Процедуры построения эталона для данного диктора;
3. Принятие решения на основе сравнения с эталонами.

2. Выделение информативных признаков

Важнейшим элементом успешного распознавания дикторов является выбор информативных признаков, способных эффективно представлять информацию об особенностях речи конкретного диктора. Требования к ним таковы:

- эффективность представления информации об особенностях речи конкретного диктора;
- простота измерения;
- стабильность во времени;
- частое и естественное появление в речи;
- практическая независимость от акустической среды;
- невосприимчивость к имитации.

Индивидуальные характеристики голоса определяются уникальностью строения артикулярного аппарата человека: строение голосовых связок, степень натяжения голосовых связок, скорость открывания и закрывания голосовой щели, объем и конфигурация речевого тракта.

Так, одним из ключевых признаков является частота основного тона F_0 — частота импульсов голосового источника, возникающая в результате колебания голосовых связок. При этом периодичность колебаний может нарушаться вследствие изменений амплитуды, частоты, фазы колебаний, наличия шума, поэтому под частотой основного тона понимают среднюю оценку на некотором интервале.

Одной из лучших характеристик гласоподобных звуков считаются формантные частоты (форманты), которые являются проявлениями резонансных частот речевого тракта диктора в акустическом сигнале [7]. Таким образом, они являются важнейшим параметром, характеризующим спектр (распределение энергии или амплитуды по частотам) речевого сигнала, которые определяют как концентрацию энергии в ограниченной частотной области. Форманта характеризуется частотой, шириной и амплитудой. За частоту форманты принимают частоту максимальной амплитуды в пределах форманты. Другими словами, форманта — это некоторый амплитудный всплеск на графике спектра, а его частота — частота пика этого всплеска (рис. 3).

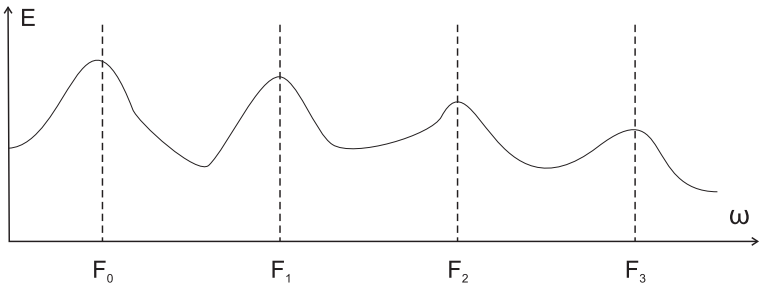

Рис. 3. Форманты речевого сигнала

Следует учитывать, что характерные признаки голоса должны вычисляться на определённых сегментах речевого сигнала. Частота основного тона — на гласоподобных участках; форма речевого тракта, которая характеризуется формантными частотами, — на гласных звуках; скорость артикуляции определяется по длительностям переходных процессов между артикуляторно-акустическими сегментами. Для выделения индивидуальных характеристик голоса целесообразно использовать только гласные и сонорные согласные звуки, хотя есть работы по идентификации на базе шипящих [3]. Таким образом, одной из основных задач является разработка надёжного алгоритма сегментации речевого сигнала и определения типа сегмента [4]. Методы сегментации в свою очередь можно условно разделить на две группы: основанные и не основанные на моделях. Методы, основанные на моделях, в основном используют кодирование по линейному предсказанию (LPC) и меры близости для оценивания спектральных изменений между последовательными кадрами речевого сигнала. При использовании модели LPC вычисляется гауссовская функция правдоподобия для обеих гипотез, и для каждого n-го кадра находится отношение правдоподобия $LR(n)$. Если на определённом кадре n_0 отношение $LR(n_0)$ превышает некий фиксированный порог, детектируется изменение. Методы, не основанные на моделях, использу-

ют различные алгоритмы обработки, такие как параметрическая фильтрация. Эти методы пытаются обнаружить спектральные изменения речевого сигнала, непосредственно рассматривая речевой спектр с использованием нескольких мер спектральных расстояний. Это означает, что если $S_1(f)$ и $S_2(f)$ — спектры двух соседних кадров сигнала, то изменение будет отмечено, если расстояние между ними превысит заданный порог [8].

3. Построение эталона

Для параметризации речевого сигнала с целью построения эталона используют две основные группы представлений — на базе преобразований Фурье и на базе линейного предсказания. Однако есть исследования альтернативных походов, на базе вейвлет-преобразований, частотных цифровых фильтров, гауссовых смесей.

Одним из способов параметризации с использованием кепстральных характеристик при построении эталона служит модель:

$$M = \{K12, \Delta K12, E, \Delta E\},$$

где $K12$ — двенадцать мел-частотных кепстральных коэффициентов (MFCC); $\Delta K12$ — двенадцать характеристик дельты MFCC; E — энергетическая характеристика; ΔE — дельта-характеристика энергии. Итого: 26 характеристик на эталон [6].

Модель также можно представить в виде набора средних значений энергии вейвлет-коэффициентов для каждого уровня детализации:

$$M = \{W_n, \Delta W_n\},$$

где W_n — значения средней энергии вейвлет-коэффициентов для десяти уровней детализации; ΔW_n — значения среднего квадратического отклонения вейвлет-коэффициентов для десяти уровней детализации; n — число уровней детализации вейвлет-преобразования. Итого: 20 характеристик на каждый эталон.

Также используются комбинации вейвлет и Фурье-анализа превосходящих по итогам экспериментов моделей на базе мел-частотных кепстральных коэффициентов и линейного предсказания [9].

4. Алгоритмы принятия решения

При принятии решения, в простейшем случае, вычисляется вероятность или расстояние от тестовых векторов информативных признаков для образца, поданного на вход системы, до эталонных векторов (моделей дикторов) и сравнивается полученное значение с порогом, часто фиксированным для всех дикторов. Могут быть использованы также механизмы нормализации для повышения устойчивости при наличии таких мешающих факторов, как: вариабельность произнесения парольной фразы одного и того же диктора, настроение диктора,

разные манеры и интонации произнесения, болезнь горла, громкость произнесения (шёпот или громкий голос) и т. д. Наиболее широко для этих целей используется когортный метод и метод мировых моделей.

Также для классификации вводимого речевого образца могут использоваться следующие методы собственно распознавания: HMM (скрытые марковские модели), моделирующие речевой сигнал на основе теоретико-вероятностных схем, DTW (динамическое программирование), базирующийся на евклидовой метрике, ANN (нейронные сети), в основе которых лежит процедура предварительного обучения, байесовский классификатор, FRIS-функции [1].

При верификации осуществляется оценка близости предъявляемого образца к эталону (модели данного диктора) и производится сравнение этой оценки с порогом, который может изменяться, чтобы осуществлять обмен между ошибками FAR и FRR. При равнозначности ошибок ($FAR = FRR = EER/2$) порог фиксирован. Критерий качества связан с ошибками 1-го и 2-го родов при проверке простой гипотезы при простой альтернативе.

При идентификации критерий качества определяется вероятностью отнесения диктора к заданной группе, правильного распознавания диктора из группы, вероятностью перепутать диктора при отнесении его к заданной группе, принять чужака за своего в группе и отождествить с конкретным диктором из группы, что связано с более общей ситуацией проверки гипотезы при сложной альтернативе. Принятие решения производится по минимуму расстояния между предъявляемым образцом и ближайшей моделью из набора моделей голосов дикторов, входящих в заданную группу. Отбор на предмет принадлежности к группе осуществляется путём сравнения указанного расстояния с порогом.

5. Проблемы голосовой идентификации

Основной проблемой для систем идентификации являются изменчивость речевого сигнала, связанная с произношением самого диктора, различия в условиях записи при регистрации пользователей и идентификации, шумы и искажения в каналах связи. Так, при хорошем отношении сигнал/шум (SNR +20 dB) достигнута точность идентификации 98 %, но уже 81 % — при +10 dB, что говорит о необходимости дальнейшего исследования в пользу способов предобработки и помехоустройчивых методов распознавания.

Таким образом, по-прежнему перспективными направлениями исследования остаются:

- поиск новых помехоустойчивых информативных признаков, связанных с характеристиками голосового источника, и формы артикуляционного тракта;
- новые решающие правила, минимизирующие ошибки 1-го и 2-го рода;
- создание полноценной речевой базы для тестирования систем идентификации с большим числом дикторов различного возраста, акцента, особенностями произношения, эмоционального состояния, записанных на различных условиях записи и с разной частотой дискретизации.

Литература

1. Борисова И. А. Алгоритм таксономии FRiS-Tax // Научный вестник НГТУ. 2007. № 3. С. 3–12.
2. Галунов В. И. Верификация и идентификация говорящего. СПб.: СПбГУ, 2002. URL: http://www.auditech.ru/article/ver_obz.doc (дата обращения: 20.10.2010).
3. Криводубский О. А., Федоров Е. Е. Моделирование особенностей речи диктора // Математичні машини і системи. 2008. № 1.
4. Леонов А. С., Макаров К. С., Сорокин В. К., Цыплихин А. К. Кодовая книга для речевых обратных задач // Информационные процессы. 2005. Т. 5. № 2. С. 101–119.
5. Мартынович П. А., Свириденко В. А. Системы верификации и идентификации диктора от SPIRIT Corp. // Доклады на конференции «BIOMETRICS 2003 AIA RUII». М., 2002.
6. Медведев М. С. Фонемная сегментация речевого сигнала с использованием вейвлет-преобразования // Доклады на конференции ИВТ СО РАН. Новосибирск, 2004.
7. Цыплихин А. И. Анализ и автоматическая сегментация речевого сигнала: дис.... канд. техн. наук. М., 2006. 149 с.
8. Li T. H., Gibson J. D. Speech Analysis and Segmentation by Parametric Filtering // IEEE Transactions on Speech and Audio Processing. 1996. Vol. 4 (3). P. 203–213.
9. Mporas J., Ganchev T. Comparison of speech features on the speech recognition task // Journal of Computer Science. 2007. Vol. 3 (8). P. 608–616.