Лилия Исааковна Цирульник, *с. н. с. лаборатории распознавания и синтеза речи Объединённого института проблем информатики Национальной академии наук Беларуси, кандидат технических наук,*

Дмитрий Александрович Покладок, *м. н. с. лаборатории распознавания и синтеза речи ОИПИ НАН Беларуси*

СИСТЕМА СИНТЕЗА РЕЧИ ПО ТЕКСТУ ДЛЯ МОБИЛЬНЫХ ТЕЛЕФОНОВ

Системы синтеза речи по тексту в настоящее время широко применяются на персональных компьютерах. Использование же синтезаторов речи по тексту на мобильных телефонах очень ограничено, поскольку последние характеризуются низким быстродействием и малым объёмом памяти, что не позволяет напрямую «перенести» на них уже существующие синтезаторы.

В настоящей статье предлагается новая архитектура системы синтеза речи по тексту, в которой обработка текста происходит на сервере, а обработка речевого сигнала — на телефоне. Описываемые алгоритмы обработки речевого сигнала имеют линейную вычислительную сложность и позволяют синтезировать речевой сигнал в реальном масштабе времени.

Введение

Системы синтеза речи по тексту к настоящему моменту достигли высокого качества как по критериям разборчивости и естественности синтезируемого голоса, так и по техническим характеристикам, что способствует их широкому применению в практических приложениях, например, в центрах обработки вызовов, при управлении сложными объектами, для создания аудио- книг и т.д. Всё более широкое распространение получает и использование систем синтеза речи на мобильных устройствах, таких как карманные персональные компьютеры или смартфоны. Это и озвучивание SMS-сообщений, и чтение писем электронной почты, и озвучивание указаний автомобильной навигационной системы. Использование систем синтеза речи на мобильных телефонах, тем не менее, ограничено. Такое ограничение связано с тем, что большинство мобильных телефонов, используемых в настоящее время, характеризуются низким быстродействием и относительно небольшим объёмом памяти. В то же время современные системы синтеза речи требуют большого объёма памяти для хранения лингвистических и акустических ресурсов, что не позволяет напрямую «перенести» существующие системы на мобильные платформы.

Существует три возможные схемы работы системы синтеза речи по тексту на мобильных телефонах:

1. Серверная, при которой система синтеза речи полностью расположена на сервере. Абоненту на мобильный телефон передаётся синтезированный речевой сигнал.

2. Клиентская, при которой система синтеза речи расположена полностью на мобильном телефоне.

3. Распределённая, при которой система синтеза речи частично расположена на сервере, частично — на мобильном телефоне.

Первая схема реализована, в частности, компанией MATERNA Information & Communications GmbH для предоставления услуги SMS2Voice (SMS2Fix) пользователям некоторых мобильных операторов в России и Украине [1]. Услуга позволяет отправлять текстовые сообщения, которые передаются синтезированным голосом на мобильные и стационарные номера.

Достоинствами данной схемы являются: возможность выбора метода синтеза, обеспечивающего наилучшее качество синтезируемой речи (поскольку нет ограничений на объём памяти и быстродействие); возможность воспользоваться данной услугой любому пользователю вне зависимости от технических характеристик его телефона; возможность модификации и обновления системы синтеза речи независимо от пользователей; высокая степень защиты системы от нелегального использования.

Очевидно, что подобным образом можно было бы передавать на мобильный телефон не только SMS-сообщения, но и любую информацию, озвученную на сервере, с использованием системы синтеза речи.

Однако данная схема имеет следующие недостатки: абонент услышит сообщение только один раз; передаваемая на мобильный телефон речевая информация имеет в несколько раз больший объём, чем исходная текстовая информация, что влечёт дополнительную существенную нагрузку на канал связи; прекращение функционирования хотя бы одного узла вызывает остановку всей службы.

Вторая из перечисленных схем получила достаточно широкое распространение. Именно по такой схеме работают программы Acapela TTS for Windows Mobile [2], Nuance TALKS [3], Mobile Speak [4] и др. В этих продуктах синтез речи по тексту полностью осуществляется на смартфонах под управлением операционных систем Windows Mobile или Symbian.

При всех очевидных преимуществах данной схемы она имеет существенный практический недостаток: смартфоны, на которых возможна работа этих систем синтеза речи, составляют только 7% рынка мобильных телефонов [5].

Третья схема до настоящего времени не была реализована, хотя она имеет высокий потенциал. Принцип работы в этой схеме основан на разделении операций между сервером и мобильным телефоном: обработка текста выполняется на сервере, в то время как работа с речевым сигналом осуществляется на мобильном телефоне. Преимущества данной схемы: возможность сохранять озвученные сообщения на телефоне; возможность выбирать просодические стили и различные голоса для синтеза; возможность использования на большинстве мобильных телефонов.

В данной работе описывается система синтеза речи по тексту для мобильных телефонов, для реализации которой выбрана последняя из описанных схем. В первом разделе представлена архитектура разработанной системы; блок обработки речевого сигнала, который работает на мобильном телефоне, описан в разделе 2; раздел 3 посвящён описанию особенностей программной реализации блока обработки речевого сигнала на языке программирования Java. Раздел 4 — заключение — суммирует основные положения данной статьи.

1. Общая структура системы синтеза речи по тексту

Система синтеза речи по тексту (рис.1) содержит два основных блока: блок преобразования текста и блок работы с речевым сигналом [6]. На первом этапе входной орфографический текст преобразуется в последовательность просодических синтагм с указанием интонационного типа каждой синтагмы, причём синтагма представлена последовательностью аллофонов (оттенков фонем в речевом потоке). На втором этапе из базы данных (БД) звуковых волн аллофонов извлекаются требуемые аллофоны, вычисляются целевые значения частоты основного тона (F_0), амплитуды (A) и длительности (T) для каждого аллофона, звуковые волны аллофонов модифицируются в соответствии с целевыми просодическими значениями и соединяются в непрерывный речевой сигнал.

Блок анализа и преобразования текста (рис. 2) содержит модули лингвистической, просодической и фонетической обработки.

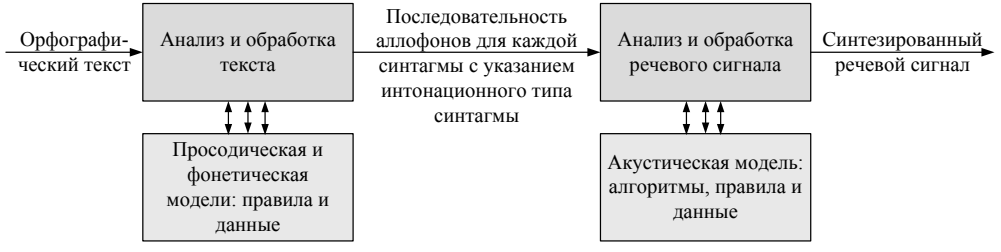

Рис. 1. Общая структурная схема системы синтеза речи по тексту

Рис. 2. Структура модуля обработки текста

Лингвистическая и просодическая обработки включают деление орфографического текста на фразы; преобразование чисел, аббревиатур, сокращений; деление фраз на просодические синтагмы; расстановку словесных ударений; деление синтагм на акцентные единицы (где под акцентной единицей понимается слово или группа слов с одним сильным ударением); маркировку интонационного типа синтагмы. Основными ресурсами лингвистического и просодического блоков являются грамматический словарь, а также правила морфологии и синтаксиса. Словарь используется для определения словесного ударения и лексико-грамматических характеристик каждого слова текста. Правила морфологии и синтаксиса используются для деления текста на фразы, фраз — на синтагмы, синтагм — на акцентные единицы, а также для определения интонационного типа синтагм.

Затем каждая интонационно размеченная синтагма поступает на фонетический процессор, который выполняет следующие задачи: фонетическое транскрибирование орфографического текста; определение позиционных и комбинаторных аллофонов; генерация аллофонных и мульти-аллофонных последовательностей, которые необходимо синтезировать.

Результат работы модуля обработки текста — последовательность синтагм с указанием интонационного типа каждой синтагмы, где каждая синтагма представлена последовательностью аллофонов — поступает в модуль обработки речевого сигнала.

В модуле обработки речевого сигнала (рис. 3) на первом этапе из речевой БД извлекаются речевые реализации аллофонов, соответствующие именам аллофонов во входной последовательности. Затем из БД просодических элементов извлекается просодический контур для соответствующего стиля и соответствующего типа синтагмы. После этого вычисляются целевые значения F_0 , A, T. Такая последовательность шагов алгоритма обусловлена тем, что вычисление целевых значений F_0 должно осуществляться для каждого периода основного тона каждого вокализованного аллофона, а число периодов основного тона в аллофонах определяется после их извлечения из речевой БД.

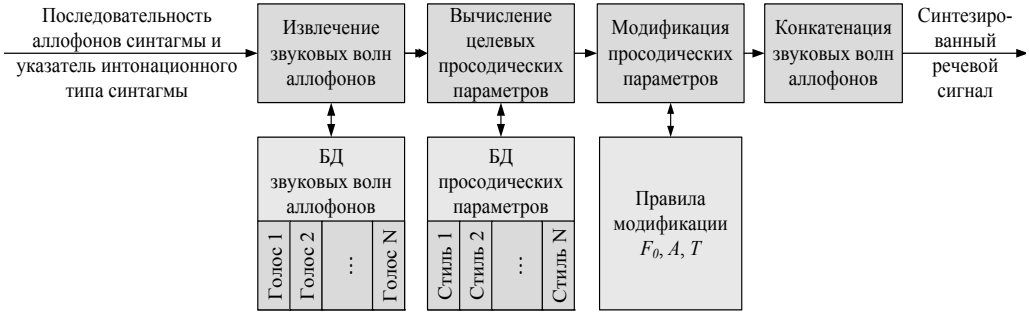

Рис. 3. Структура модуля обработки речевого сигнала

Необходимо отметить, что модуль обработки текста требует гораздо большего объёма памяти для хранения и использования ресурсов, чем модуль обработки речевого сигнала, а также характеризуется большей вычислительной сложностью. Действительно, один из основных лингвистических ресурсов — грамматический словарь русского языка — содержит более 3,5 миллиона словоформ [7]. Учитывая, что русский является флективным языком, целесообразно хранить словарь в виде компактной двухуровневой структуры, в которой первый уровень содержит неизменные части слов, а второй уровень — флексии. Для хранения в таком виде словаря объёмом 3,5 миллиона словоформ требуется порядка 50 мБ дискового пространства. Вычислительная сложность операций поиска слова в словаре равна $O(\log_2 n)$, где n — количество слов в словаре. Вычислительная сложность всех операций, выполняемых лингвистическим, просодическим и фонетическим процессорами текста, включая операции поиска слова в словаре, равна $O(m)*O(n)$, где m — число слов входного текста.

Ресурсы блока обработки речевого сигнала — БД звуковых волн аллофонов и БД просодических параметров — требуют соответственно 750 кБ для одного голоса и 11 кБ для одного интонационного стиля. Вычислительная сложность алгоритмов обработки речевого сигнала равна $O(k)$, где k — количество аллофонов во входной последовательности.

При оценке алгоритмов синтеза речи по тексту важно учитывать тактовую частоту устройства, на котором должна быть реализована система, поскольку среднее время обработки одной синтагмы должно быть намного меньше, чем время воспроизведения синтезированной синтагмы, которое составляет в среднем от 1 до 10 секунд. Время обработки одной синтагмы на персональном компьютере с тактовой частотой 1,3 ГГц составляет 0,4–0,5 секунды.

Большинство современных мобильных телефонов обладает следующими характеристиками: доступная память от 128 КБ до 4 МБ, 32-битный RISC-процессор с тактовой частотой от 50МГц и выше, поддержка языка программирования Java ME и конфигурации CLDC. Такие характеристики не могут обеспечить достаточно быструю работу блока обработки текста, но являются удовлетворительными для быстрой работы блока обработки речевого сигнала.

Таким образом, оптимальной для реализации на большинстве современных мобильных телефонов является архитектура, при которой блок обработки текста расположен на сервере, в то время как блок обработки речевого сигнала находится на мобильном телефоне. Дополнительным достоинством такой архитектуры является возможность синтеза речи по одному и тому же размеченному тексту, поступающему с сервера, с использованием различных голосов и различных просодических стилей, находящихся на мобильном телефоне.

2. Обработка речевого сигнала на мобильном телефоне

Из четырёх блоков обработки речевого сигнала, представленных на рис. 3, наибольший интерес представляют блоки вычисления целевых просодических пара-

метров и модификации просодических параметров в речевом сигнале. Особенности работы этих блоков описаны в данном разделе.

2.1. Блок вычисления целевых просодических параметров

Для вычисления целевых просодических параметров используется просодическая модель Портретов Акцентных Единиц (ПАЕ-модель) [7]. Согласно ПАЕ-модели, каждое предложение состоит из последовательности синтагм, где под синтагмой понимается самостоятельная в интонационном смысле часть предложения. Каждая синтагма, в свою очередь, состоит из одной или более акцентных единиц. Акцентная единица (АЕ) является минимальной просодической единицей и состоит из одного или более слов, имеющих лишь один полноударный гласный. Интонационно значимыми элементами АЕ являются ядро (полноударный гласный), предъядро (все фонемы, предшествующие полноударному гласному), и заядро (все фонемы, следующие за полноударным гласным).

Основное предположение ПАЕ-модели в том, что топологические свойства просодических параметров не зависят от конкретного фонетического контекста и количества слогов в предъядре и заядре для конкретного типа интонации. Таким образом, просодические характеристики могут задаваться «портретами» акцентных единиц, которые указывают нормированные значения F_0, A и T на участках предъядра, ядра и заядра.

Полный набор таких «портретов», содержащий интонационные характеристики для разных типов синтагм, составляет просодический стиль. БД просодических параметров, используемая на данном этапе, может содержать несколько различных просодических стилей.

В блок вычисления целевых просодических параметров информация подаётся по синтагмам. На первом этапе определяется просодический тип синтагмы и количество АЕ в ней, после чего из БД просодических параметров извлекается соответствующий просодический «портрет». Затем в каждой АЕ выделяются аллофоны, составляющие предъядро, ядро и заядро.

Для каждого аллофона на основе ритмического «портрета», а также на основе положения аллофона в предъядре, ядре или заядре вычисляется коэффициент изменения длительности аллофона (в процентах) k_a. Затем вычисляется целевое значение длительности каждого i-того аллофона T'_{ai}:

$$T'_{ai} = \frac{T_{ai} \cdot k_a}{100}, \qquad (1)$$

где T_{ai} — исходная длительность аллофона.

Вычисление целевых интонационных значений осуществляется только для вокализованных аллофонов, при этом интонационные характеристики вычисляются (в отличие от ритмических характеристик) не для всего аллофона, а для каждого периода основного тона аллофона. На основе интонационного «портрета», а также на основе положения аллофона в предъядре, ядре или заядре вычисляются нормализованные целевые значения F_0. Затем с учётом диапазона частоты основного тона используемой рече-

вой БД вычисляются целевые значения длительностей периодов:

$$T'_{0i} = \frac{f_{дискр} \cdot 100}{F_{0norm\,i} \cdot (F_{0\max} - F_{0\min}) + F_{0\min} \cdot 100}, \quad (2)$$

где T'_{0i} — целевое значение i-того периода основного тона (количество отсчётов сигнала);

$f_{дискр}$ — частота дискретизации сигнала;

$F_{0norm\,i}$ — нормализованное (в диапазоне [0..100]) значение частоты основного тона *i*-того периода;

F_{0max}, F_{0min} — максимальное и минимальное значение частоты основного тона для речевой базы.

Полученные целевые значения передаются в блок модификации просодических параметров в речевом сигнале.

2.2. Блок модификации просодических параметров в речевом сигнале

Модификация просодических параметров в речевом сигнале осуществляется с использованием метода «плавной сшивки» периодов основного тона [6]. Основным достоинством данного метода является неизменность речевого сигнала на участке периода основного тона, который соответствует моменту схлопывания голосовых связок, что позволяет сохранить индивидуальные тембральные характеристики обрабатываемого голоса. Несомненным достоинством алгоритма «плавной сшивки», важным при его реализации на мобильных телефонах, является линейная вычислительная сложность.

Процесс уменьшения периода показан на рис. 4 и 5. Удаляется часть периода длиной *N,* где

$$N = T_0 - T'_{0i}; \quad (3)$$

T_0 — текущая длина *i*-того периода, T'_0 — целевая длина периода основного тона.

Удаляемая часть смещается и накладывается на предшествующую часть периода (рис.4). Накладывание двух участков сигнала происходит путём плавного уменьшения первого сигнала и увеличения второго сигнала (рис. 5).

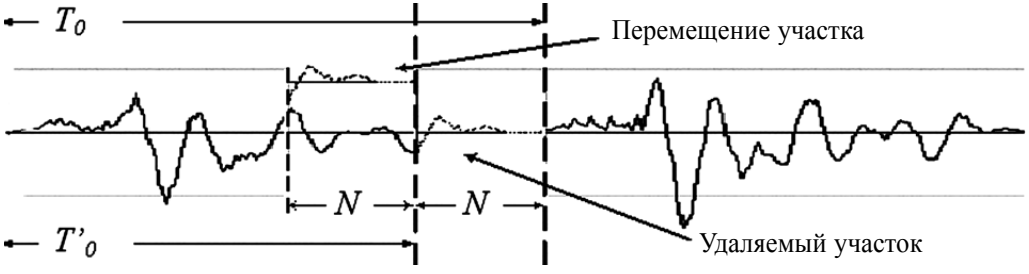

Рис. 4. Перемещение удаляемого участка сигнала

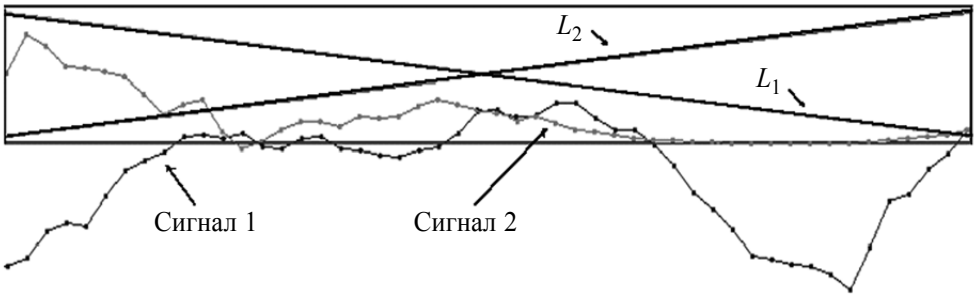

Рис. 5. Формирование переходного участка путём «плавной сшивки» двух сигналов

Модификация сигнала при уменьшении длительности периода основного тона осуществляется в соответствии с формулой:

$$\tilde{s}(n) = \frac{(N-n)\cdot s(n) + n \cdot s(n+N)}{N},$$

$$(T_0^{'} - N) \le n \le T_0^{'}, \qquad (4)$$

где $\tilde{s}(n)$ — результирующий речевой сигнал;

$s(n)$ — исходный сигнал.

Аналогичная процедура осуществляется при увеличении периода основного тона [6]. При этом результирующий речевой сигнал $s(n)$ вычисляется в соответствии с формулой:

$$\tilde{s}(n) = \frac{(T_0^{'} - n)\cdot s(n) + n \cdot s(n-N)}{T_0^{'}},$$

$$N \le n \le T_0^{'}. \qquad (5)$$

3. Программная реализация системы на мобильном телефоне

Блок обработки речевого сигнала реализован на языке Java Mobile Edition [9] для минимальной конфигурации CLDC 1.0 [10] и профиля MIDP 2.0 [11], что позволяет использовать его практически на любом современном мобильном телефоне. В следующих разделах описывается пользовательский интерфейс созданной системы и особенности её программной реализации.

3.1. Пользовательский интерфейс системы

Главное меню системы (рис. 6а) включает выбор текстового файла для воспроизведения, просмотр/изменение настроек, непосредственно воспроизведение и справочную информацию, которая содержится в элементе меню «О программе». Настройки включают выбор голосовой базы и выбор просодического стиля для синтеза речи (рис. 6 б). Для синтеза речи пользователь должен сначала указать текстовый файл, содержащий размеченный текст, затем выбрать элемент меню «воспроизведение». При воспроизведении (рис. 6в) в системе реализованы функции паузы/возобновления, а также остановки воспроизведения.

Рис. 6. Интерфейс системы: а) главное меню; б) выбор голосовой БД и просодического стиля для синтеза речи; в) воспроизведение речевого сигнала

3.2. Особенности программной реализации системы

Звуковые волны аллофонов, содержащиеся в речевой БД, хранятся в формате WAVE PCM. В процессе генерации речевого сигнала они извлекаются из БД, модифицируются в соответствии с описанными выше алгоритмами и помещаются в буфер для воспроизведения. После того, как очередная речевая синтагма подготовлена, она воспроизводится с использованием стандартного класса J2ME Player.

Поскольку процесс генерации речевого сигнала должен происходить практически одновременно с процессом воспроизведения синтезированной речи, в системе реализована многопоточность. При этом главный поток управляет действиями двух дочерних, один из которых генерирует очередную речевую синтагму, другой — воспроизводит. Первый из потоков характеризуется высокой трудоёмкостью выполнения и поэтому имеет больший приоритет, чем второй поток. В то же время выходные данные первого потока являются входными данными второго, поэтому потоки синхронизированы с тем, чтобы работа второго потока всегда начиналась после завершения первого. Такая синхронизация осуществляется главным потоком.

ЗАКЛЮЧЕНИЕ

Разработанная система была успешно протестирована на мобильных телефонах Motorola, Sony-Ericsson, LG, которые характеризуются тактовой частотой ARM-процессора от 68 до 115 МГц, объемом памяти от 3 500 до 4 200 КБ, поддержкой конфигурации CLDC 1.0 и профиля MIDP 2.0. Система позволяет синтезировать речевой сигнал в реальном времени на мобильных телефонах с ARM-процессорами седьмого поколения.

Экспертная оценка качества синтезированной речи показала, что оно не уступает качеству синтезированной речи, получаемому на персональных компьютерах с использованием тех же методов обработки текстовой и речевой информации.

Созданная система является универсальной в том смысле, что замена используемой голосовой базы (например, мужской на женскую) не требует дополнительной обработки входного текста.

Созданная система может быть модернизирована с целью озвучивания входящих SMS-сообщений и электронных текстов, полученных через сеть Internet.

ЛИТЕРАТУРА

1. SMS2Voice. Сервис голосовых сообщений [Электронный ресурс]. — Электронные данные. — Режим доступа: http://voice.s-soft.org. — Дата доступа: 10.12.09.

2. Acapela TTS for Windows Mobile [Электронный ресурс]. — Электронные данные. — Режим доступа: http://www.acapela-group.com/acapela-tts-for-windows-mobile-2-2-speech-solutions-tts.html. — Дата доступа: 01.06.10.

3. Nuance TALKS [Электронный ресурс]. — Электронные данные. — Режим доступа: http://www.nuance.com/talks/. — Дата доступа: 10.12.09.

4. Mobile Speak [Электронный ресурс]. — Электронные данные. — Режим доступа: http://www.codefactory.es/en/products.asp?id=316. — Дата доступа: 10.12.09.

5. Gartner Says Worldwide Mobile Phone Sales Grew 17 Per Cent in First Quarter 2010. — Press Release. — Электронный ресурс. — Режим доступа: http://www.gartner.com/it/page.jsp?id=1372013– Дата доступа: 01.07.10.

6. *Лобанов Б.М., Цирульник Л.И.* Компьютерный синтез и клонирование речи Мн.: Белорусская наука, 2008. 344 с.: ил.

7. *Жадинец Д.В., Сизонов О.Г., Цирульник Л.И.* Электронные словари русского и белорусского языков для двуязычной системы синтеза речи по тексту // Танаевские чтения: доклады тр. межд. конф., Минск, 28 марта 2007 г. М.: Объединённый институт проблем информатики, 2007. С. 65–69.

8. *Lobanov B., Karnevskaya E.* Auditory Estimation of Effectiveness of the AUP-Stylization Model of the Melodic Contour TTS-synthesis and Voice Cloning. — Proc. 13-th Int. Conf. SPECOM'2009, June 21–25, 2009, St.-Pet.. P. 130–135.

9. Java ME at a Glance [Электронный ресурс]. — Электронные данные. — Режим доступа: http://www.oracle.com/technetwork/java/javame/overview/index.html. — Дата доступа: 1.08.10.

10. Connected Limited Device Configuration (CLDC); JSR 30, JSR 139 Overview [Электронный ресурс]. — Электронные данные. — Режим доступа: http://www.oracle.com/technetwork/java/overview-142076.html. — Дата доступа: 1.08.10.

11. Mobile Information Device Profile (MIDP); JSR 37, JSR 118 Overview [Электронный ресурс]. — Электронные данные. — Режим доступа: http://www.oracle.com/technetwork/java/overview-140208.html. — Дата доступа: 1.08.10.