УДК 519.688

ТЕХНОЛОГИЯ СИНТЕЗА ЕСТЕСТВЕННОЙ РЕЧИ С ИСПОЛЬЗОВАНИЕМ БАЗЫ ДАННЫХ НЕБОЛЬШОГО ОБЪЕМА

П.Г. Чистиков[a], А.О. Таланов[b], Д.С. Захаров[a, b], А.И. Соломенник[c]

[a] Университет ИТМО, Санкт-Петербург, Россия, chistikov@speechpro.com

[b] ООО «ЦРТ», Санкт-Петербург, Россия

[c] ООО «Речевые технологии», Минск, Беларусь

Аннотация. Представлен подход к созданию голоса для системы синтеза естественной речи в условиях малого объема исходного речевого материала. Эффективное решение данной проблемы необходимо для задачи восстановления голоса (синтез потерянных фрагментов записи на основе доступного материала известного диктора, например актера). Представленная система синтеза речи является гибридной, так как комбинирует достоинства систем, основанных на скрытых марковских моделях и методе Unit Selection. Подход, описанный в работе, использует статистические модели интонационных параметров, что позволяет сохранять в синтезированной речи особенности произношения диктора. Описан процесс подготовки базы данных для синтеза, в том числе и решение проблемы нехватки исходного речевого материала для обучения модели. Специальные алгоритмы конкатенации и модификации звуковых элементов помогают корректировать их параметры в соответствии с требованиями, обеспечивают общую тональную гладкость и уменьшают искажения в спектральной области на границах объединяемых фрагментов. Аудитивные тесты показали эффективность предложенных решений и доказали, что синтез естественной речи возможен даже в условиях малой речевой базы (вплоть до одного часа речи).

Ключевые слова: синтез речи, восстановление голоса, скрытые марковские модели, метод Unit Selection, модификация речи.

Благодарности. Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

METHOD OF HIGH-QUALITY SPEECH SYNTHESIS WITH A SMALL DATABASE USAGE

P.G. Chistikov[a], A.O. Talanov[b], D.S. Zakharov[a, b], A.I. Solomennik[c]

[a] ITMO University, Saint Petersburg, Russia, chistikov@speechpro.com

[b] Speech Technology Center Ltd., Saint Petersburg, Russia

[b] Speech Technology Ltd., Minsk, Belarus

Abstract. We propose an approach to synthesizing high-quality speech in view of a small initial speech database. A robust method for solving this problem is vital for voice restoration (recovery of the lost fragments of recordings based on available speech material of a well-known person, e.g. an actor). The proposed TTS (text-to-speech) system is a hybrid one that combines the advantages of both HMM- and Unit Selection-based TTS systems. The paper deals with the approach based on statistical models of intonation parameters, which makes it possible to preserve the speaker's pronunciation in synthesized speech. We describe the preparation of the database and the solution to the problem of shortage of original speech material for model training. Special algorithms of speech element concatenation and modification are effective to correct parameters according to the requirements, provide overall tonal smoothness and reduce spectral distortion at the boundaries of concatenated elements. Listening tests showed the efficiency of the proposed methods and proved the possibility of high-quality speech synthesis even with a small speech database (right up to one hour of speech).

Keywords: speech synthesis, voice restoration, hidden Markov models, Unit Selection, speech modification.

Acknowledgements. This work was partially financially supported by the Government of the Russian Federation, Grant 074-U01.

Введение

В последнее время, благодаря активному проведению исследований, технология синтеза речи серьезно продвинулась вперед. В результате синтезированная речь сейчас воспринимается естественно, и мы можем услышать ее во многих местах. Самым популярным подходом для достижения высокого качества синтезированной речи является метод Unit Selection [1, 2]. Системы синтеза речи (Text to Speech – TTS), основанные на скрытых марковских моделях (СММ), звучат хуже, более «механически» из-за посторонних призвуков, таких как жужжание, звон [3]. Но платой за хорошее качество является необходимость наличия большой речевой базы (примерно 10 часов звукозаписей) [2, 4]. Усугубляет положение то, что каждый звуковой файл в базе данных должен быть размечен с высокой точностью. Это увеличивает стоимость и временные затраты на ее подготовку [2]. TTS-системы могут быть использованы для контакт-центров, чтения аудиокниг, систем голосовой помощи. В данных областях не принципиально, чей голос используется. Более приоритетными являются качество синтезированной речи и общие впечатления от голоса.

Несмотря на вышесказанное, существуют некоторые сферы, где очень важна специфичность голоса. В качестве примеров можно назвать клонирование голоса и восстановление голоса, т.е. воссоздание потерянных фрагментов записи на основе доступного речевого материала хорошо известной личности (актера, диктора). Сложность этой задачи заключается, главным образом, в небольшом количестве дос-

тупного речевого материала. К тому же существующие записи диктора зачастую акустически различны: они могут быть сделаны с разных микрофонов, в совершенно различных условиях и в течение большого периода времени. По этой причине удается накопить лишь небольшую речевую базу с материалом достаточно хорошего для TTS качества.

Существует ряд исследований, посвященных синтезу речи в условиях недостаточного количества речевого материала [5–8]. Все они основаны на речевой базе неопределенного диктора и последующем получении целевого голоса с помощью адаптационных техник, приложенных к элементам речи или акустическим моделям, в качестве которых обычно используются CMM. Оба этих подхода генерируют недостаточно естественную речь из-за дополнительных эффектов, возникающих в результате применения адаптации в акустической области. Для решения этой проблемы предложена гибридная TTS-система [9], в которой интонация целевого диктора моделируется на основе речевой базы другого диктора путем применения методов адаптации. Синтез речи выполняется по речевой базе целевого диктора с применением методов Unit Selection и с использованием специальных методов модификации и конкатенации элементов речи. Это позволяет реализовывать синтез естественной речи даже с использованием малой речевой базы (вплоть до одного часа речи), что подтверждается результатами аудитивных тестов.

Описание гибридной системы синтеза речи

Структурно TTS можно разделить на две части: часть подготовки речевой базы и часть непосредственно синтеза речи (рис. 1). Основное назначение первой состоит в подготовке базы речевых элементов и создании модели целевого голоса на базе параметров речи диктора-донора.

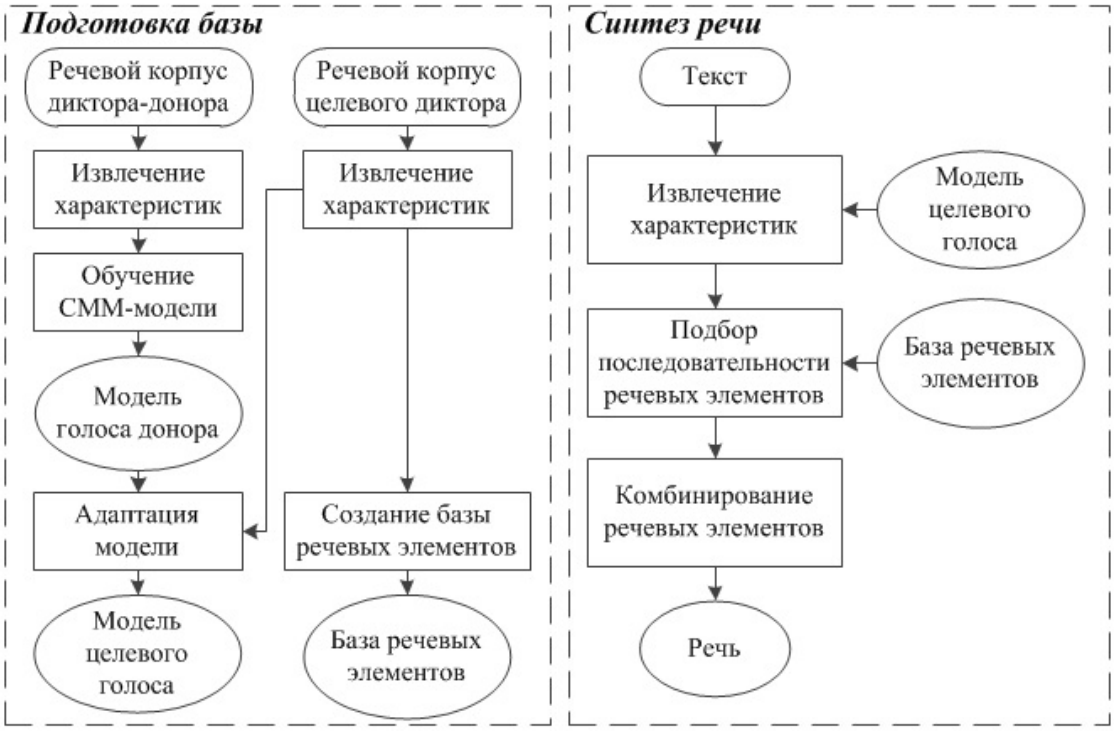


Рис. 1. Структура гибридной системы синтеза речи

Исходными данными для этапа подготовки служат два речевых корпуса (размеченные речевые базы данных): корпус диктора-донора, содержащий 8–10 часов речи, для обучения интонационной модели и корпус целевого диктора для подготовки базы речевых элементов. Каждый из них содержит набор звуковых файлов (каждый файл – одно записанное предложение) и набор меток для каждого файла. Эти метки содержат информацию об элементах речи [10–12], их лингвистические и акустические характеристики [9–14], вычисленные для двух корпусов. В первую очередь происходит обучение речевой модели целевого диктора. Эта модель является набором CMM, которые генерируют параметры речевых элементов для различного контекста. В качестве параметров используются мел-частотные кепстральные коэффициенты (Mel Frequency Cepstral Coefficients – MFCC), основной тон (ОТ), энергия и длительность сигнала. Более подробно создание модели голоса целевого диктора описывается в следующем разделе. Одновременно с обучением из речевого материала целевого диктора создается речевая база для синтеза. Она содержит проиндексированный набор элементов для выполнения быстрого поиска по необходимым признакам: имя фонемы, имена соседних фонем, MFCC на границах элементов, энергия, ОТ, длительность.

Синтез речи происходит путем выбора из речевой базы элементов в соответствии с моделью голоса диктора и последующего их объединения. На вход системы подается текст без какой-либо предварительной обработки. Из него формируется последовательность аллофонов, а также лингвистические и просодические параметры для каждого из них. С использованием этой информации и моделей голоса для каждого аллофона генерируются акустические параметры – MFCC, ОТ, энергия и длительность. Далее на основании этих параметров по методу Unit Selection [4] из базы выбирается последовательность наиболее подходящих звуковых элементов.

Подобранные элементы должны быть модифицированы в соответствии с предсказанными параметрами и объединены в единый звуковой поток. Эти финальные шаги очень важны с учетом малого объема материала целевого диктора. В таких условиях наиболее вероятна ситуация, когда невозможно найти в базе речевой элемент, соответствующий моделируемым параметрам. Для обеспечения нужной интонации и общей гладкости синтезированной речи необходимо применение специальных подходов, описанных ниже в разделе «Модификации речевых элементов».

Моделирование интонации

Моделирование интонационных параметров начинается с извлечения ряда характеристик из звукового файла. Для этого исходный сигнал разбивается на короткие участки (кадры) длительностью по 25 мс. Извлекаются следующие характеристики:

– последовательность $\{C_1,...,C_K\}$ MFCC-векторов [15]. Каждый вектор состоит из 25 коэффициентов и характеризует спектральное представление кадра. K – общее число кадров;
– последовательность $\{F0_1,...,F0_K\}$ значений ОТ;
– последовательность $\{E_1,...,E_K\}$ значений энергии.

После этого вычисляются лингвистические и просодические характеристики каждого аллофона для всех файлов обучающей базы данных [9, 14].

На следующем шаге создается СММ-прототип каждого аллофона речевой базы донора. Каждая СММ – однонаправленная (слева направо) без ветвлений, с числом состояний $N = 5$. Каждый выход наблюдаемого вектора $\mathbf{o}^i$ для i-го кадра содержит 5 потоков $\mathbf{o}^i = [\mathbf{o}_1^{iT}, \mathbf{o}_2^{iT}, \mathbf{o}_3^{iT}, \mathbf{o}_4^{iT}, \mathbf{o}_5^{iT}]^T$. Их содержимое продемонстрировано на рис. 2. Поток 1 содержит вектор MFCC-коэффициентов, их первую и вторую производные (Δ MFCC и Δ^2 MFCC на рисунке). Поток 2 содержит вектор значений ОТ (F0), поток 3 содержит значения первой производной ОТ (Δ F0), поток 4 – второй производной ОТ (Δ^2 F0). Поток 5 содержит вектор значений энергии кадра En, его первой и второй производных (Δ En и Δ^2 En на рис. 2).

Для каждой k-й СММ длительность нахождения в N состояниях представляется в виде вектора $\mathbf{d}^k = [d_1^k,...,d_N^k]^T$, где d_n^k – длительность нахождения в n-ом состоянии. Кроме того, вектор длительности моделируется N-мерным гауссовым распределением. Далее выходные вероятности для состояний векторов длительностей переоцениваются с помощью алгоритма Баума–Уолша, тем же способом, что и вероятности параметров речи [15].

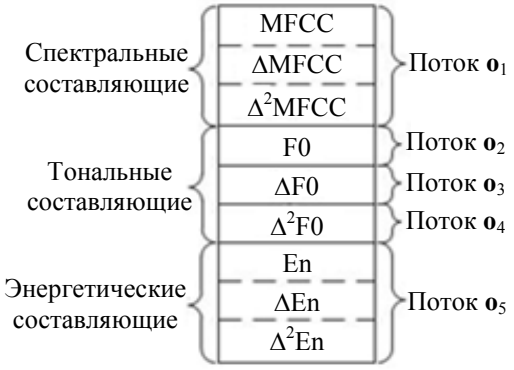


Рис. 2. Структура вектора наблюдения

На заключительном шаге интонационная модель диктора-донора адаптируется таким образом, чтобы быть максимально приближенной к параметрам целевого диктора. Адаптация выполняется с использованием процедуры, описанной в [16]. Во время построения модели к состояниям СММ для значений MFCC, F0, энергии и их производных применяется кластеризация, основанная на деревьях. Это же делается и для состояний модели длительности. При завершении процесса генерируется $5N+1$ различных акустических деревьев решений: N – для MFCC-компонент и их первых и вторых производных, $3N$ – для значений ОТ, N – для характеристик энергии и одно дерево для состояний длительности. Выполне-

ние этого этапа дает возможность генерировать речевые параметры для элементов, отсутствующих в базе данных, и тем самым обеспечивать вывод необходимых параметров даже в условиях недостаточности обучающих данных. В итоге мы имеем интонационную модель целевого диктора, которая в дальнейшем будет использоваться для предсказания речевых параметров синтезируемых предложений.

Модификация речевых элементов

После того как из базы выбраны подходящие элементы, может потребоваться их модификация по длительности или ОТ согласно параметрам, предсказанным интонационной моделью. Этот шаг необходим для обеспечения свойственной диктору интонации в синтезируемом предложении. В описываемой системе используется модель линейного предсказания (Linear Prediction – LP) [17] для извлечения ошибки предсказания $\mathbf{e}[n]$, ее модификации с помощью метода TD-PSOLA [17] и восстановления исходного сигнала с новым значением ОТ из модифицированной ошибки $\mathbf{e}'[n]$.

Вычисление $\mathbf{e}[n]$ происходит согласно следующему выражению:

$$\mathbf{e}[n] = \mathbf{s}[n] - \mathbf{a}^T \cdot \mathbf{s}'[n-1],$$

где $\mathbf{s}'[n-1] = \left[\mathbf{s}[n-1], \mathbf{s}[n-2], ..., \mathbf{s}[n-P]\right]^T$, $\mathbf{a} = [a_1, a_2, ..., a_P,]^T$, $P = 25$.

Вектор LP-коэффициентов вычисляется согласно выражению

$$\mathbf{a}_n = \mathbf{R}^{-1}[n-1] \cdot \mathbf{p}[n], \quad \mathbf{R}^{-1}[n-1] = \sum_{i=0}^{n} \mathbf{s}'[i-1] \cdot \mathbf{s}'^T[i-1], \quad \mathbf{p}[n] = \sum_{i=0}^{n} \mathbf{s}'[i-1] \cdot \mathbf{s}[i].$$

Значения $\mathbf{p}[n]$ могут быть вычислены рекурсивно, чтобы избежать больших вычислительных затрат: $\mathbf{p}[n] = \mathbf{s}[n-1] \cdot \mathbf{s}[n] + \mathbf{p}[n-1]$.

С использованием LP-коэффициентов, полученных в процессе анализа, LP-модель может применяться в процессе синтеза модифицированного сигнала $\mathbf{s}_m[n]$ с желаемой огибающей ОТ следующим образом: $\mathbf{s}_m[n] = \mathbf{e}'[n] + \mathbf{a}^T \cdot \mathbf{s}_m[n-1]$. LP-модель определяется для каждого отсчета, что приводит к сглаживанию переходов между соседними моделями.

При условии, что границы периода ОТ $\mathbf{p}[n]$ определены метками $\mathbf{p}_m[n]$, огибающая ОТ может быть модифицирована желаемым образом. Для этого новые позиции меток $\mathbf{p}'_m[n]$ определяются в соответствии с новыми длительностями периодов ОТ $\mathbf{p}'[n]$, где $\mathbf{p}'[n] = \beta[n] \cdot \mathbf{p}[n]$, а $\beta[n]$ – коэффициент модификации периода ОТ. Значение этого коэффициента определяется в соответствии с требуемой просодической составляющей или при общей автоматической коррекции. Новая позиция метки ОТ $\mathbf{p}'_m[n]$ устанавливается путем вставки периода $\mathbf{p}'[n]$ между двумя последовательными метками. Таким образом, метка будет расположена на позиции $n + \mathbf{p}'[n]$, где n – предыдущая метка ОТ. На следующем шаге осуществляется связывание каждой новой метки $\mathbf{p}'_m[n]$ с ближайшим соответствующим пиком в оригинальном сигнале $\mathbf{p}_m[n]$. Это выполняется напрямую сравнением временных индексов $\mathbf{p}_m[n]$ и $\mathbf{p}'_m[n]$.

На последнем шаге генерации нового сигнала каждый пик в оригинальном сигнале взвешивается двумя полуокнами Хэннинга от предыдущей метки ОТ до текущей. Полученные сегменты складываются с перекрытием, в соответствии с рассчитанным ранее $\mathbf{p}'[n]$.

Конкатенация речевых элементов

На последнем шаге, когда элементы подобраны и откорректированы под желаемые интонационные параметры, они объединяются в единый звуковой поток. При этом решается проблема несовпадения спектра и ОТ на границах элементов. Для ее решения применяются алгоритмы оптимального объединения и сглаживания ОТ. Оптимальное объединение происходит путем коррекции границ элементов для минимизации спектрального искажения в позиции конкатенации. Этот процесс продемонстрирован на рис. 3. Пусть дифоны a1_a2 и a2_a3 были выбраны в качестве оптимальных. Оригинальная граница дифонов в исходном файле находится в точке *B*. Эта позиция сравнивается с соседними двумя – *A* и *C*, которые образованы сдвигом *B* на смещение Δ, обычно равное двум или трем периодам ОТ. Оптимальная позиция объединения речевых элементов P_{opt} рассчитывается согласно следующей формуле: $P_{opt} = \underset{P=\{A,B,C\}}{\arg\min} L_2(P)$, где

$$L_2(P) = \sqrt{\sum_{i=1}^{M} \left(c_{Li}(P) - c_{Ri}(P)\right)^2}. \quad (1)$$

В формуле (1) $c_{Li}(P)$ – это *i*-й MFCC-коэффициент на границе P для левого дифона, а $c_{Ri}(P)$ – соответственно *i*-й MFCC-коэффициент на границе P для правого дифона. В данном случае используется 12 MFCC-коэффициентов.

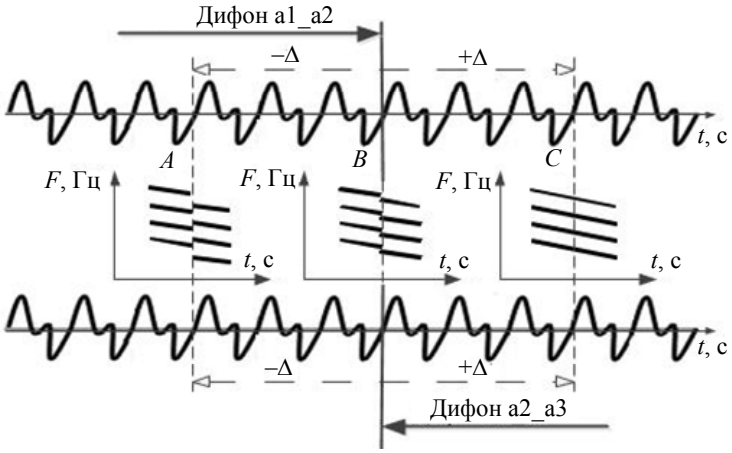


Рис. 3. Корректировка границ перед оптимальным объединением

Идея сглаживания основного тона на границе элементов состоит в том, чтобы, с одной стороны, устранить скачок огибающей ОТ, а с другой – сохранить локальные отклонения значений ОТ, чтобы речь не получилась слишком статичной.

Пусть $\mathbf{p}_L = \{p_{L1}, p_{L2}, ..., p_{LN}\}$ – N значений длительностей периодов ОТ для левого элемента, а $\mathbf{p}_R = \{p_{R1}, p_{R2}, ..., p_{RM}\}$ – M значений длительностей периодов ОТ для правого элемента. Вместе они образуют вектор огибающей ОТ $\mathbf{p} = \{p_{L1}, p_{L2}, ..., p_{LN}, p_{R1}, p_{R2}, ..., p_{RM}\}$, который должен быть сглажен. Результирующая огибающая $\mathbf{p}' = \{p'_1, p'_2, ..., p'_{N+M}\}$ может быть рассчитана следующим образом.

В первую очередь огибающая $\mathbf{p}$ разделяется на две составляющие – низкочастотная составляющая $\mathbf{p}_m$ и высокочастотная $\mathbf{p}_f$, где $\mathbf{p}_m[i] = \alpha \cdot \mathbf{p}[i] + (1-\alpha) \cdot \mathbf{p}_m[i-1]$, а $\mathbf{p}_f = \mathbf{p} - \mathbf{p}_m$, $0 < \alpha < 1$. Затем $\mathbf{p}_m$ сглаживается методом построения кривых Безье:

$$\mathbf{p}'_m[i] = \sum_{j=1}^{N+M} \mathbf{p}_m[j] \cdot b_{j-1,N+M-1}\left(\frac{i-1}{N+M-1}\right),$$

где $b_{i,n}(t) = C_i^n \cdot t^i \cdot (1-t)^{n-i}$, $C_i^n = \dfrac{n!}{i! \cdot (n-i)!}$. Результирующий вектор вычисляется как сумма сглаженной и высокочастотной составляющих: $\mathbf{p}' = \mathbf{p}'_m + \mathbf{p}'_f$.

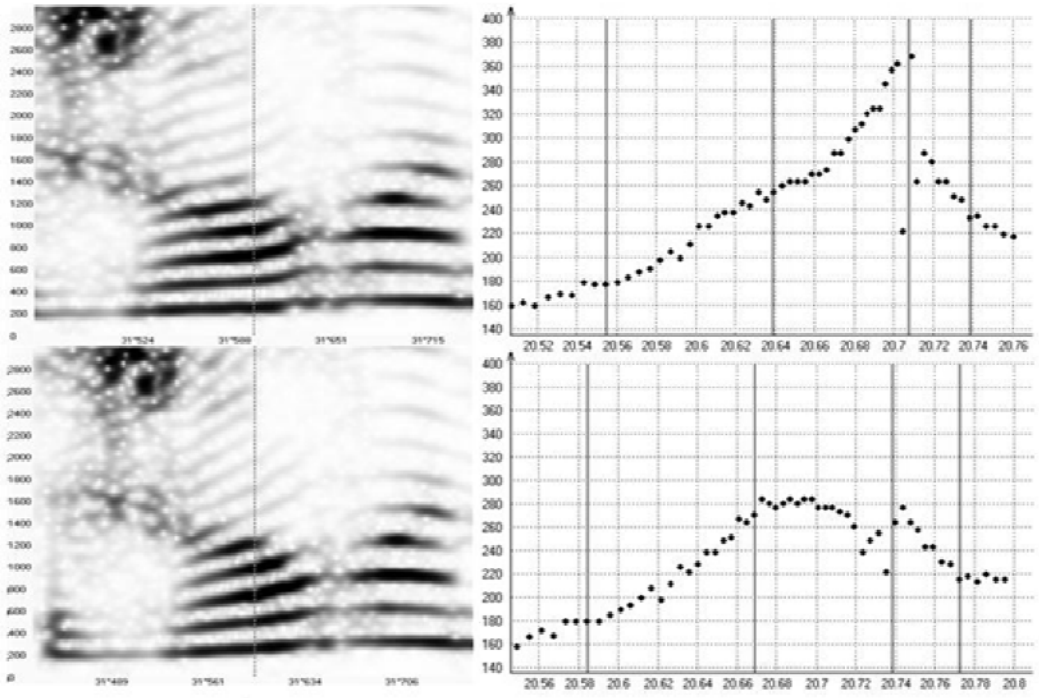
Экспериментальные результаты

В данном разделе описываются эксперименты, выполненные с использованием алгоритмов оптимального объединения и сглаживания ОТ на границах элементов.

Рис. 4, а, иллюстрирует результат применения техники оптимального объединения элементов. Само место объединения отмечено вертикальной линией. Диаграмма вверху показывает спектр при объединении на изначально выбранной позиции элементов, нижний – после применения техники оптимального объединения. Отметим, что во втором случае спектр выглядит более естественно. На верхнем рисунке имеется обрыв спектральных пиков, чего нет на нижнем рисунке.

Результат сглаживания ОТ можно увидеть на рис. 4, б. Верхняя часть диаграммы показывает оригинальную огибающую ОТ, нижняя часть – модифицированную. Разрыв ОТ (отмеченный вертикальной линией) был сглажен с сохранением небольших отклонений на всей огибающей.

Чтобы оценить предложенную систему для различного размера речевых баз была проведена экспертная оценка MOS (Mean Opinion Score). Рассматривалось две составляющие: естественность и разборчивость речи. Критерии оценивания стандартны [18] и представлены в таблице. Результаты данной процедуры показаны на рис. 5.

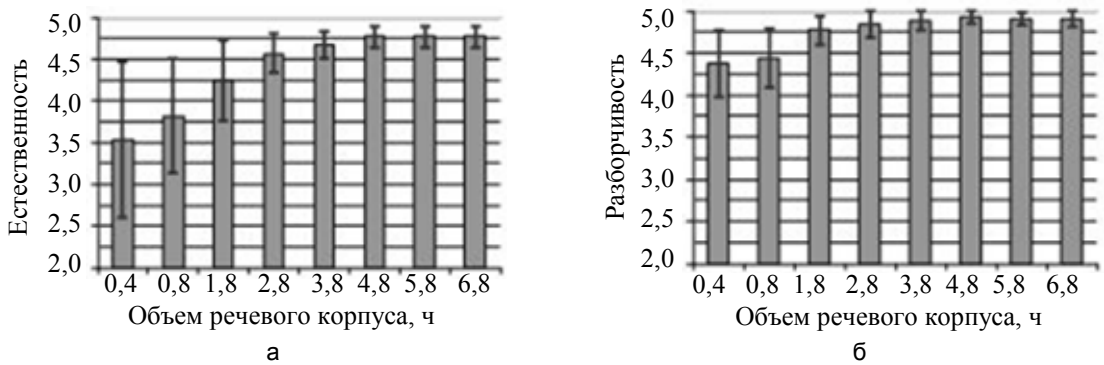


а б

Рис. 4. Результат работы алгоритмов конкатенации элементов: фрагмент спектра на позиции конкатенации: сверху – без корректировки границы, снизу – с корректировкой (а); огибающая ОТ: сверху – без сглаживания, снизу – со сглаживанием (б)

Критерий естественности речи	Оценка	Критерий разборчивости речи	Оценка
Естественно звучащая речь, возможны небольшие искажения. Отсутствует хрип, треск. Высокая разборчивость	> 4,5	Полностью разборчивая речь	5
Небольшие недостатки естественности и разборчивости, слабое влияние одного типа искажений (шум, звон, хрип, др.)	3,6–4,5	Разборчивая речь, понимание без лишних усилий	4,6–4,9
Заметные нарушения естественности и разборчивости. Наличие нескольких типов искажений (шум, звон, хрип, др.)	2,6–3,5	Разборчивая речь, понимание с небольшими усилиями	3,6–4,5
Наличие постоянных искажений (шум, звон, хрип, др.). Значительные нарушения естественности и разборчивости	1,7–2,5	Почти разборчивая речь, понимание с усилиями	2,6–3,5
Сильные искажения: шум, звон, хрип, треск, механическое звучание голоса. Значительные нарушения естественности и разборчивости	< 1,7	Частично разборчивая речь, понимание со значительными усилиями	< 2,5

Таблица. Критерии оценки естественности и разборчивости речи



а б

Рис. 5. Зависимость экспертной оценки естественности (а) и разборчивости (б) речи от объема базы

Заключение

В работе представлено описание системы синтеза естественной речи с использованием речевой базы небольшого объема. Основной целью являлось решение проблемы восстановления голоса, а также создание голоса в условиях сильного ограничения речевого материала. Задача была реализована с применением гибридного подхода (CMM-синтез плюс Unit Selection). Интонация целевого диктора моделируется по речевой базе другого диктора, к которой применялись технологии адаптации речевых характеристик, а выбираемые в процессе синтеза элементы речи изменялись в соответствии с предсказанными параметрами с применением специальных техник модификации и конкатенации. Экспериментальные результаты и субъективные экспертные оценки показали эффективность этого подхода и возможность синтеза высококачественной естественной речи даже в условиях малого количества исходных записей речи. Более того, предложенный подход уменьшает требования к предварительной разметке речевой базы и позволяет улучшить качество синтезированной речи.

Литература

1. Breuer S., Bergmann S., Dragon R., Möller S. Set-up of a unit-selection synthesis with a prominent voice // Proc. 5th International conference on Language Resources and Evaluation. Genoa, 2006. P. 293–296.
2. Matoušek J., Tihelka D., Šmídl L. On the impact of annotation errors on unit-selection speech synthesis // Lecture Notes in Computer Science. 2012. V. 7499. P. 456–463.
3. Yamagishi J., Zen H., Toda T., Tokuda K. Speaker-independent HMM-based speech synthesis system – HTS-2007 system for the blizzard challenge 2007 // Proc. Blizzard Challenge-2007. Bonn, Germany, 2007. P. 1–6.
4. Hunt A.J., Black A.W. Unit selection in a concatenative speech synthesis using a large speech database // Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 96. Atlanta, USA, 1996. V. 1. P. 373–376.
5. Phung T.-N., Mai C.L., Akagi M. A concatenative speech synthesis for monosyllabic languages with limited data // Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2012. Hollywood, US, 2012. P. 1–10.
6. Meng F., Wu Z., Meng H., Jia J., Cai L. Hierarchical english emphatic speech synthesis based on HMM with limited training data // Proc. 13th Annual Conference of the International Speech Communication Association, InterSpeech 2012. Portland, US, 2012. V. 1. P. 466–469.
7. Tsuzuki R., Zen H., Tokuda K., Kitamura T., Bulut M., Narayanan S. Constructing emotional speech synthesizers with limited speech database // Proc. INTERSPEECH 2004-ICSLP. Jeju Island, Korea, 2004. P. 1185–1188.
8. Phung T. N., Luong M. C., Akagi M. A hybrid TTS between unit selection and HMM-based TTS under limited data conditions // Proc. 8th ISCA Speech Synthesis Workshop. Barcelona, Spain, 2013. P. 279–284.
9. Chistikov P.G., Korolkov E.A., Talanov A.O. Combining HMM and unit selection technologies to increase naturalness of synthesized speech // Компьютерная лингвистика и интеллектуальные технологии. 2013. № 12-2. С. 2–10.
10. Чистиков П.Г., Корольков Е.А., Таланов А.О., Соломенник А.И. Гибридная технология синтеза речи на основе скрытых марковских моделей и алгоритма Unit Selection // Изв. вузов. Приборостроение. 2013. Т. 56. № 2. С. 33–38.
11. Соломенник А.И., Таланов А.О., Соломенник М.В., Хомицевич О.Г., Чистиков П.Г. Оценки качества синтезированной речи: проблемы и решения // Изв. вузов. Приборостроение. 2013. Т. 56. № 2. С. 38–42.
12. Чистиков П.Г., Хомицевич О.Г., Рыбин С.В. Статистические методы автоматического определения мест и длительности пауз в системах синтеза речи // Изв. вузов. Приборостроение. 2014. Т. 57. № 2. С. 28–32.
13. Chistikov P.G., Korolkov E.A. Data-driven speech parameter generation for Russian text-to-speech system // Компьютерная лингвистика и интеллектуальные технологии. 2012. № 11. С. 103–111.
14. Chistikov P., Khomitsevich O. Improving prosodic break detection in a Russian TTS system // Proc. of the 15th International Conference on Speech and Computer, SPECOM 2013. Pilsen, Czech Republic, 2013. V. 8113. P. 181–188.
15. Zen H., Tokuda K., Masuko T., Kobayashi T., Kitamura T. A hidden semi-Markov model-based speech synthesis // IEICE Transactions on Information and Systems. 2007. V. E90-D. P. 825–834.
16. Yamagishi J., Kobayashi T. Adaptive training for hidden semi-Markov model // Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'05. Philadelphia, US, 2005. V. 1. Art. N 1415126. P. I365–I368.
17. Taylor P. Text-to-Speech Synthesis. Cambridge University Press, 2009. 626 p.
18. ГОСТ Р 50840-95. Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. Введ. 01.01.1997. М: Издательство стандартов, 1996. 234 с.

Чистиков Павел Геннадьевич – кандидат технических наук, научный сотрудник, Университет ИТМО, Санкт-Петербург, Россия, chistikov@speechpro.com

Таланов Андрей Олегович – кандидат технических наук, руководитель отдела синтеза речи, ООО «ЦРТ», Санкт-Петербург, Россия, andre@speechpro.com

Захаров Дмитрий Сергеевич – студент, Университет ИТМО, Санкт-Петербург, Россия; младший научный сотрудник, ООО «ЦРТ», Санкт-Петербург, Россия, zakharov-d@speechpro.com

Соломенник Анна Ивановна – научный сотрудник, ООО «Речевые технологии», Минск, Беларусь, solomennik-a@speechpro.com

Pavel G. Chistikov – PhD, scientific researcher, ITMO University, Saint Petersburg, Russia, chistikov@speechpro.com

Andrey O. Talanov – PhD, Head of the speech synthesis department, Speech Technology Center Ltd., Saint Petersburg, Russia, andre@speechpro.com

Dmitry S. Zakharov – student, ITMO University, Saint Petersburg, Russia; Junior researcher, Speech Technology Center Ltd., Saint Petersburg, Russia, zakharov-d@speechpro.com

Anna I. Solomennik – scientific researcher, Speech Technology Ltd., Minsk, Belarus, solomennik-a@speechpro.com

Принято к печати 12.05.14
Accepted 12.05.14