УДК 519.688

П. Г. Чистиков, Е. А. Корольков, А. О. Таланов, А. И. Соломенник

# ГИБРИДНАЯ ТЕХНОЛОГИЯ СИНТЕЗА РЕЧИ НА ОСНОВЕ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ И АЛГОРИТМА UNIT SELECTION

Рассматриваются особенности построения системы синтеза русской речи с использованием двух наиболее распространенных подходов — статистического, на основе скрытых марковских моделей, и конкатенативного, на основе алгоритма Unit Selection. Для решения задачи моделирования интонации разработана методика создания модели голоса русскоязычного диктора. Эксперименты показывают повышение естественности звучания синтезируемой речи.

***Ключевые слова:*** *синтез речи, скрытые марковские модели, Unit Selection, модель голоса.*

**Введение.** Синтез речи по тексту представляет собой автоматический перевод последовательности символов произвольного текста в соответствующую им последовательность отсчетов звукового сигнала [1—3]. Существует несколько подходов к организации автоматического синтеза речи по тексту. К основным можно отнести синтез по правилам (формантный синтез), артикуляторный синтез, компилятивный, синтез на основании статистических моделей [4—8].

Наиболее распространены в настоящее время подходы, основанные на алгоритме Unit Selection (US) и на скрытых марковских моделях (СММ-синтез). Первый позволяет достичь максимальной естественности звучания синтезированной речи при использовании корректно отсегментированной на разных уровнях сбалансированной речевой базы данных большого объема. В то же время статистический подход, обеспечивая меньшую естественность звучания синтезированной речи (эффект роботизированности), обладает следующими преимуществами:

1) позволяет легко модифицировать характеристики голоса с помощью адаптации/интерполяции моделей дикторов, в то время как алгоритм US позволяет получить речь, стиль которой не отличается от стиля речевой базы;

2) звучание речи, полученной на основе СММ-технологии, естественно, однако в ней отсутствуют резкие, не обусловленные контекстом перепады по частоте и энергии, обычно присущие конкатенативному синтезу. Кроме того, при применении алгоритма US результат синтеза может существенно ухудшиться в случае отсутствия подходящего звукового элемента в базе данных. При использовании моделей отсутствующие в обучающей выборке звуковые элементы синтезируются на основе средних значений, максимально приближенных к требуемым, благодаря применению технологии кластеризации контекстов, основанной на деревьях. Это позволяет добиться разборчивости синтезированной речи в условиях ограниченного количества звуковых единиц в различных контекстах;

3) позволяет разрабатывать новый голос за гораздо меньшее время, а также требует значительно меньше памяти для хранения речевой базы.

В предлагаемой гибридной системе используются оба подхода: оптимальная последовательность звуковых элементов подбирается из речевого корпуса диктора по классическому алгоритму Unit Selection, но с применением статистической интонационной модели, обученной на той же базе, что позволяет повысить естественность звучания синтезируемой речи по сравнению с реализацией на US или только на основе СММ-технологии.

**Описание системы.** Функционально и структурно систему можно разделить на подсистемы подготовки звуковой базы данных (подготовительный этап) и синтеза речи (рис. 1). Звуковая база данных строится на основе речевого корпуса, состоящего из совокупности звуковых

файлов, каждый из которых содержит запись одного предложения, и соответствующего ему набора файлов разметки, содержащих необходимую информацию о представленных в предложении звуковых единицах [9—12]. По файлам разметки строится индексная база, обеспечивающая быстрый поиск по целевым характеристикам, таким как имя аллофона, имена аллофонов слева и справа, коэффициенты MFCC (Mel-Frequency Cepstral Coefficients) на границах аллофона, энергия на границах, частота основного тона на границах и длительность аллофона.

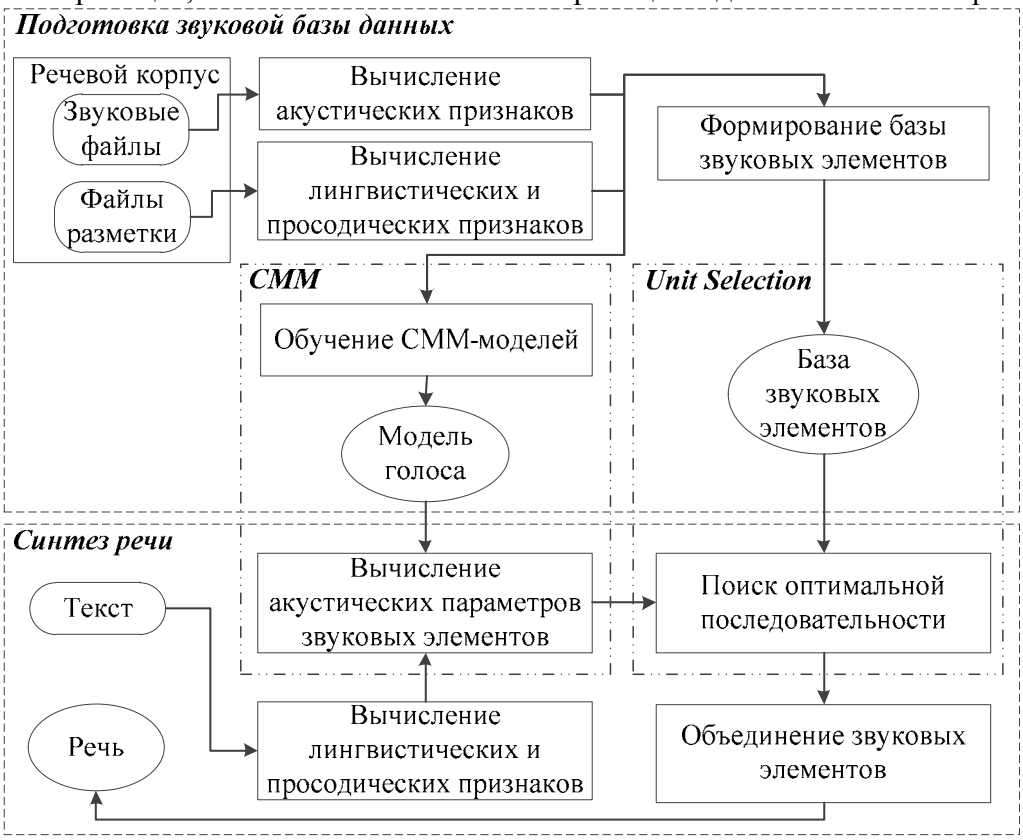


*Рис. 1*

Процедура моделирования параметров голоса начинается с расчета набора характеристик для всех звуковых файлов [13, 14]. Каждый такой набор описывает короткий участок сигнала (кадр) длительностью 25 мс. В качестве характеристик используются следующие параметры.

— Последовательность $\{c_1, \ldots, c_K\}$ векторов MFCC коэффициентов [15], каждый вектор состоит из 25 коэффициентов и характеризует спектральную огибающую сигнала на фрейме; $K$ — общее количество фреймов.

— Последовательность $\{F0_1, ..., F0_K\}$ значений частоты основного тона.

На следующем шаге для каждого аллофона на основе файлов разметки вычисляется набор лингвистических и просодических признаков, включающий в себя 7 аллофонных (имена стоящего перед предыдущим, предыдущего, текущего, следующего и следующего за следующим аллофонов; позиция от начала и конца слога), 13 слоговых, 8 словных и 3 синтагматических признака.

Далее для каждого аллофона создаются прототипы CMM-моделей. Каждая модель имеет $N$ состояний, допустимы переходы в себя или в следующее состояние. В предлагаемой системе $N = 5$. Каждый выходной вектор наблюдений $\overline{\mathbf{o}}^i$ состоит из четырех потоков $\overline{\mathbf{o}}^i = [\overline{\mathbf{o}}_1^{iT}, \overline{\mathbf{o}}_2^{iT}, \overline{\mathbf{o}}_3^{iT}, \overline{\mathbf{o}}_4^{iT}]^T$: первый содержит значения MFCC, их первых и вторых производных, второй — значение частоты основного тона (ЧОТ), третий — значение первой производной, четвертый — второй производной ЧОТ.

Плотность вектора наблюдений $\bar{\mathbf{o}}^i$ на выходе из состояния $n$ СММ-модели задается следующим выражением:

$$\beta_n(\bar{\mathbf{o}}^i) = \prod_{j=1}^{4}\left[\sum_{l=1}^{R_j}\omega_{njl}\mathcal{N}^{\circ}(\bar{\mathbf{o}}_j^i;\mu_{njl},\Sigma_{njl})\right],$$

где $\mathcal{N}^{\circ}(\cdot;\mu,\Sigma)$ — плотность нормального распределения с вектором средних значений $\mu$ и матрицей ковариации $\Sigma$, $\omega_{njl}$ — весовой коэффициент $l$-й компоненты смеси $j$-го потока выходного вектора $n$-го состояния, $R_j$ — количество компонент в смеси $j$-го потока. Для $k$-й СММ-модели вектор длительности состояний $\bar{d}_k = [\bar{d}_{k1},...,\bar{d}_{kN}]^T$ моделируется $N$-мерным однокомпонентным гауссовым распределением. Вероятности выходных значений моделей спектральных параметров (MFCC и ЧОТ) и длительностей переоцениваются при помощи алгоритма Баума—Велша [16].

Процесс построения модели голоса завершается кластеризацией состояний СММ-моделей на основе деревьев решений. На данном шаге генерируются параметры отсутствующих в обучающей речевой базе элементов, что, в свою очередь, обеспечивает синтез разборчивой речи даже при небольшом объеме обучающего материала. Итоговая интонационная модель голоса состоит из $N$+1 деревьев: $N$ — для хранения по каждому из состояний параметров СММ-модели ЧОТ вместе с первой и второй производными, и одно — для параметров СММ-модели длительностей.

На вход системе синтеза подается текст без какой-либо предварительной ручной обработки. На основе текстовой информации для каждого предложения формируется целевая последовательность аллофонов и вычисляются лингвистические и просодические признаки для каждого из них. Тип и структура признаков аналогичны тем, что используются на этапе подготовки звуковой базы данных. На основе этой информации по модели голоса определяются акустические признаки каждого аллофона: значения ЧОТ, энергии и длительности. По рассчитанным акустическим и лингвистическим характеристикам из речевой базы выбирается группа наиболее подходящих звуковых элементов. Для того чтобы определить, насколько тот или иной элемент базы подходит для синтеза данной единицы, вводятся понятия стоимости замены (target cost) и стоимости связи (concatenation cost) [17].

Стоимость замены для элемента из базы $u_i$ по отношению к искомому элементу $t_i$ вычисляется по формуле:

$$C_t(u_i,t_i) = \sum_{k=1}^{p} w_{tk} C_{tk}(u_i,t_i),$$

где $C_{tk}$ — расстояние между $k$-ми характеристиками элементов, $w_{tk}$ — вес для $k$-й характеристики. Другими словами, стоимость замены есть взвешенная сумма различий в признаках между целевым (требуемым) элементом и конкретным элементом речевой базы. В качестве признаков могут выступать ЧОТ, длительность, контекст, позиция элемента в слоге, слове, количество ударных слогов во фразе и др.

Выбранные элементы должны не только мало отличаться от целевых, но и хорошо соединяться друг с другом. Стоимость связи двух элементов может быть определена как взвешенная сумма различий в признаках между двумя последовательно выбранными элементами:

$$C_c(u_{i-1},u_i) = \sum_{k=1}^{q} w_{ck} C_{ck}(u_{i-1},u_i),$$

где $C_{ck}$ — расстояние между $k$-ми характеристиками элементов, $w_{ck}$ — вес для $k$-й характеристики.

Общая стоимость целой последовательности из $n$ элементов есть сумма введенных выше стоимостей:

$$C(u,t)=\sum_{i=1}^{n}C_t(u_i,t_i)+\sum_{i=2}^{n}C_c(u_{i-1},u_i). \qquad (1)$$

Задача алгоритма Unit Selection — выбрать такое множество, которое бы минимизировало общую стоимость согласно формуле (1).

В завершение происходит объединение выбранной последовательности элементов в единый звуковой поток, на выходе представляющий собой синтезированную речь.

**Экспериментальные результаты.** Примеры работы системы представлены на рис. 2—4, на которых приведены соответственно осциллограммы, спектрограммы и графики динамики частоты основного тона для фразы „Это очень важно!“. На приведенных рисунках в верхней части представлены данные для фразы, записанной реальным диктором, а в нижней — для ее синтезированного варианта. Следует отметить, что синтезируемая фраза не была включена в обучающую выборку.

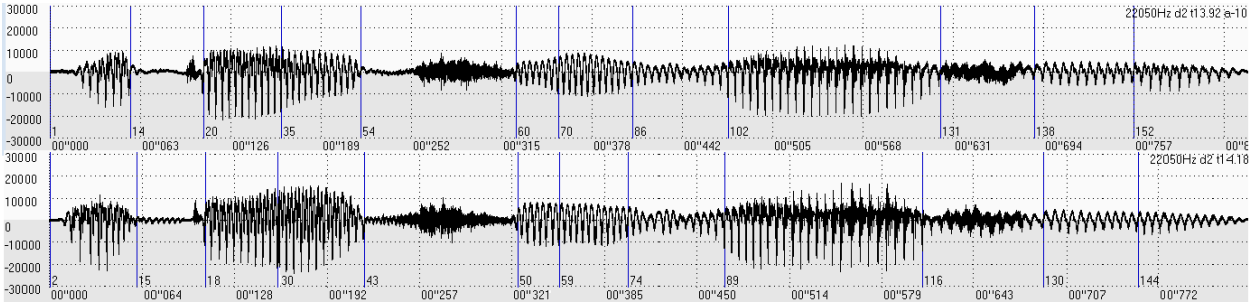

*Рис. 2*

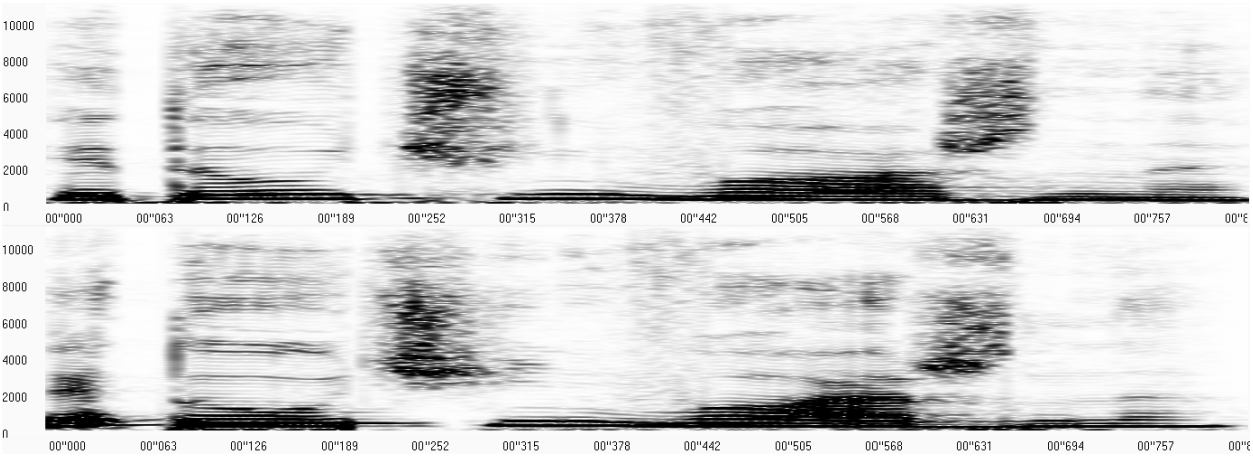

*Рис. 3*

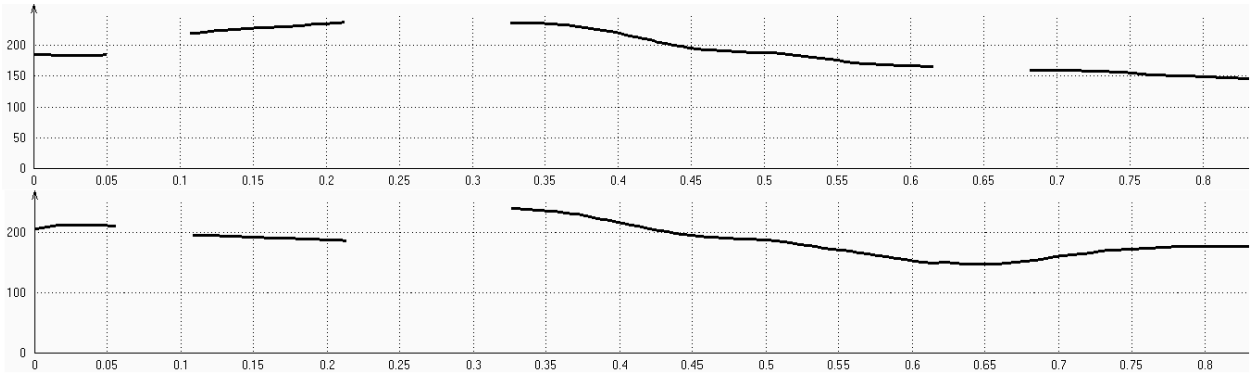

*Рис. 4*

На основе данных диаграмм можно сделать вывод, что синтезированная фраза имеет практически такие же темпоритмические и спектральные характеристики, как и ее эквивалент, произнесенный диктором, что достигается за счет определения значений этих характеристик на основе скрытых марковских моделей.

Ниже приведены результаты сравнения показателей естественности речи (значения в интервале от 0 до 5, где 5 — максимальная оценка естественности) представленной в работе системы с системой на основе метода US, лежащей в основе гибридного подхода. Пять экспериментов оценивали два (мужской и женский) голоса, данные в таблице усреднены. Как видно из результатов эксперимента, применение гибридного подхода позволило улучшить показатели естественности синтезированной речи.

| Тип подхода к синтезу | | |
|---|---|---|
| Unit Selection | гибридный подход | естественная речь |
| 4,0 | 4,3 | 4,8 |

**Заключение.** В ходе проведенных исследований была разработана гибридная система синтеза русской речи по тексту, в основе которой лежат скрытые марковские модели и алгоритм Unit Selection. Результаты испытаний показали, что по показателям естественности звучания данная система является одной из лучших среди систем синтеза на русском языке.

## СПИСОК ЛИТЕРАТУРЫ


1. *Dines J.* Model based trainable speech synthesis and its applications. Ph. D. Thesis. Brisbame, Australia: Queensland University of Technology, 2003.
2. *Dutoit Th.* Introduction au traitement de la parole // Faculte Polytechnique de Mons. 2002.
3. *Stilianou Y.* Harmonic plus noise models for speech, combined with statisticals methods, for speech and speaker modification. Ph. D. Thesis. Paris, France: Ecole Nationale Superieure des Telecommunications, 1996.
4. *Klatt D. H.* Review of text-to-speech conversion for English // J. of the Acoustical Society of America. 1987. Vol. 82. P. 737—793.
5. *Tokuda K.* HMM-based Speech Synthesis System (HTS). 2011 [Электронный ресурс]: <http://hts.sp.nitech.ac.jp>.
6. *Huang X., Acero A., Adcock J., Goldsmith J., Liu J. W.* A Trainable Text-to-Speech System // Proc. of the Intern. Conf. on Spoken Language Processing. Philadelphia, PA. 1996. Vol. 4. P. 2387—2390.
7. *Donovan R. E., Eide E. M.* The IBM Trainable Speech Synthsis System // Proc. ICSLP'98. Sydney, Australia, 1998.
8. *Donovan R. E., Ittycheriah A., Franz M., Ramabhadran B., Eide E., Viswanathan M., Bakis R., Hamza W.* Current Status of the IBM Trainable Speech Synthesis System // Proc. 4th ESCA Tutorial and ResearchWorkshop on Speech Synthesis. Scotland, UK. 2001.
9. *Продан А., Чистиков П., Таланов А.* Система подготовки нового голоса для системы синтеза "VITALVOICE" // Компьютерная лингвистика и интеллектуальные технологии. 2010. № 9 (16). С. 394—399.
10. *Смирнова Н., Чистиков П.* Программа анализа фонетических статистик в текстах на русском языке и ее использование для решения прикладных задачах в области речевых технологий // Там же. 2011. № 10 (17). С. 632—643.
11. *Чистиков П., Хомицевич О.* Автоматическое определение границ предложений в потоковом режиме в системе распознавания русской речи // Вестн. МГТУ им. Н. Э. Баумана. 2011. Вып. S. С. 117—125.
12. *Chistikov P., Khomitsevich O.* On-line automatic sentence boundary detection in a Russian ASR system // SPECOM 2011 Intern. Conf. 2011. P. 112—117.
13. *Чистиков П.Г.* Моделирование параметров русской речи в системе синтеза // Сб. тез. докл. конгресса молодых ученых. Вып. 2. СПб: НИУ ИТМО. 2012. С. 227—228.
14. *Chistikov P., Korolkov E.* Data-driven Speech Parameter Generation for Russian Text-to-Speech System // Компьютерная лингвистика и интеллектуальные технологии. 2012. № 11 (18). С. 103—111.

15. *Fukada T., Tokuda K., Kobayashi T., Imai S.* An adaptive algorithm for mel-cepstral analysis of speech // Proc. of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). 1992. P. 137—140.

16. *Zen H., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Hidden semi-Markov model based speech synthesis // Proc. of the Intern. Conf. on Spoken Language Processing (ICSLP). 2004. P. 1393—1396.

17. *Black A.W., Hunt A.J.* Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database // Proc. of ICASSP 96. Atlanta, Georgia, 1996. Vol. 1. P. 373—376.

***Сведения об авторах***

***Павел Геннадьевич Чистиков*** — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: chistikov@speechpro.com

***Евгений Александрович Корольков*** — ООО „ЦРТ“, Санкт-Петербург; научный сотрудник; E-mail: korolkov@speechpro.com

***Андрей Олегович Таланов*** — канд. техн. наук; ООО „ЦРТ“, Санкт-Петербург; руководитель отдела синтеза речи; E-mail: andre@speechpro.com

***Анна Ивановна Соломенник*** — ООО „Речевые технологии“, Минск; научный сотрудник; E-mail: solomennik-a@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

УДК 81'322.6

А. И. Соломенник, А. О. Таланов, М. В. Соломенник, О. Г. Хомицевич, П. Г. Чистиков

# ОЦЕНКА КАЧЕСТВА СИНТЕЗИРОВАННОЙ РЕЧИ: ПРОБЛЕМЫ И РЕШЕНИЯ

Рассмотрены различные аспекты проблемы оценки результатов работы систем синтеза речи. Приведен краткий обзор существующих методик оценки качества.

***Ключевые слова:*** *синтез речи, качество синтезированной речи, сравнение систем синтеза речи.*

**Введение.** Синтезированная речь в последние годы все больше используется в различных сферах, например, в банковских системах голосового самообслуживания, транспортных компаний, при проведении телефонных опросов. Синтезированными голосами „говорят“ мобильные устройства, озвучиваются аудиокниги. Поэтому задача оценки качества синтезированной речи и сравнения систем синтеза между собой становится как никогда актуальной. Однако в этой области существует немало проблем. Основной можно назвать субъективность оценок: кто-то обращает внимание на тембр синтезированного голоса или на ошибки в произношении, для кого-то голос слишком „роботизирован“ или, наоборот, „излишне живой“ и непредсказуемый. В настоящей статье будут рассмотрены существующие подходы к объективной оценке качества синтеза в целом и его отдельных компонентов.

**Оценка качества лингвистической обработки.** Системы синтеза могут сравниваться и оцениваться по следующим объективным параметрам, отражающим решение задач лингвистической обработки в системах синтеза:

1) выделение предложений в тексте и разбиение их на отдельные слова; разметка текста на буквы, специальные символы, цифры и знаки пунктуации;