УДК 004.934

СИСТЕМА КОНВЕРСИИ ГОЛОСА В РЕАЛЬНОМ МАСШТАБЕ ВРЕМЕНИ С ТЕКСТОНЕЗАВИСИМЫМ ОБУЧЕНИЕМ НА ОСНОВЕ ГИБРИДНОГО ПАРАМЕТРИЧЕСКОГО ОПИСАНИЯ РЕЧЕВЫХ СИГНАЛОВ

Азаров И.С., *к.т.н., доцент кафедры электронных вычислительных средств БГУИР, e-mail: azarov@bsuir.by*
Петровский А.А., *д.т.н., профессор, зав. кафедрой электронных вычислительных средств БГУИР, e-mail: palex@bsuir.by*

Предлагается способ конверсии голоса на основе гибридной модели параметрического описания речевого сигнала. Поиск функции конверсии выполняется с использованием образцов речи исходного и целевого дикторов с произвольным текстовым содержанием. Эффективность предложенного способа оценивается путем экспериментальной реализации системы конверсии.

Ключевые слова: конверсия голоса, параметрическое описание речевого сигнала.

Введение

Задача конверсии голоса привлекает все больше внимания со стороны пользователей и разработчиков современных систем мультимедиа. Под конверсией голоса понимается такое преобразование входного речевого сигнала, которое «переозвучивает» его голосом другого (целевого) диктора. Конверсия голоса может служить незаменимым инструментом для многих приложений в области речевых технологий. Например, используя систему конверсии, многоголосое озвучивание фильмов и радиопередач может осуществляться всего лишь несколькими актерами. Причем голоса персонажей могут быть синтезированы без участия самих целевых дикторов, поскольку в качестве эталонов могут быть использованы записи их голосов. Таким же способом может быть осуществлена реставрация старых звукозаписей и звуковых дорожек кинокартин. В таких задачах, как синтез речи по тексту, конверсия голоса также может быть полезной, так как с ее помощью возможно повышение натуральности звучания и устранение «компьютерного» акцента [1]. При применении эффективных алгоритмов конверсии возможно создание многоголосого синтезатора, используя акустическую базу одного диктора.

Несмотря на то, что предлагаемые решения становятся все более и более сложными, исследования в данной области находятся на начальной стадии, поскольку в настоящее время не существует системы конверсии, обеспечивающей абсолютную узнаваемость и разборчивость выходной речи. Ниже перечислены основные факторы, существенно затрудняющие использование систем конверсии голоса:

– сложности, связанные с параметрическим описанием речевого сигнала;

– подготовка обучающих последовательностей и поиск соответствия;

– изменчивость голосовых параметров в зависимости от большого числа факторов (экспрессия, состояние голосового тракта и т.д.).

Человеческая речь является продуктом различных психических и физиологических процессов, которые на сегодняшний день не имеют точного математического описания. Речевой сигнал имеет изменчивую структуру, комбинируя периодические и стохастические составляющие, что требует специфических средств анализа и синтеза. Это вынуждает исследователей использовать специальные средства выделения компонент различной природы и выполнять их раздельную обработку.

Преобразование параметров речи выполняется при помощи функции конверсии, которая является уникальной для заданной пары исходного и целевого дикторов. В большинстве существующих на сегодняшний день систем применяется текстозависимый способ обучения, т.е. для поиска функции конверсии используются продолжительные, фонетически сбалансированные обучающие последовательности, строго синхронизированные по времени [2-3]. Такой подход обеспечивает максимально достижимое на сегодняшний день качество конверсии, однако требует от пользователя значительных усилий, поскольку для каждого целевого и исходного диктора необходимо заново проводить весь процесс обучения. Качество конверсии сильно страдает, если тренировочные последовательности синхронизированы недостаточно хорошо. Нахождение взаимно однозначного соответствия между фрагментами речи исходного и целевого дикторов практически невозможно, поскольку одна и та же фраза произносится по-разному, в зависимости от экспрессии, настроения и интонации.

Альтернативой изложенному выше подходу является текстонезависимое обучение, предполагающее использование одной (целевой) обучающей последовательности, причем она не обязательно должна быть фонетически сбалансирована. Несмотря на то, что в данном случае часто не может быть обеспечено необходимое качество [4-6], текстонезависимый подход является наиболее перспективным из-за простоты в использовании и широкой области возможного применения.

Задача текстонезависимой конверсии голоса на сегодняшний день не имеет классических решений, поскольку относится к сложной области исследования и требует применения новых, оригинальных способов обработки речевой информации. Тем не менее, можно обозначить несколько возможных направлений, в кото-

рых следует вести исследования: классификация (кластеризация) речевой базы по акустическим и фонетическим признакам, поиск и использование унифицированных речевых параметров, создание кодовой книги конверсии с мультиголосовым базисом.

В данной работе предлагается метод конверсии голоса, использующий текстонезависимое обучение функции конверсии. Приводятся полученные практические результаты.

Система конверсии голоса с текстонезависимым обучением

Анализ современной литературы, касающейся вопроса конверсии голоса [3], [5-7], показывает, что подавляющее большинство систем конверсии с текстонезависимым обучением используют скрытые марковские модели (СММ) для поиска соответствия между фонетическими единицами исходного и целевого дикторов. Общая структура системы конверсии может быть изображена следующим образом (см. рис.1).

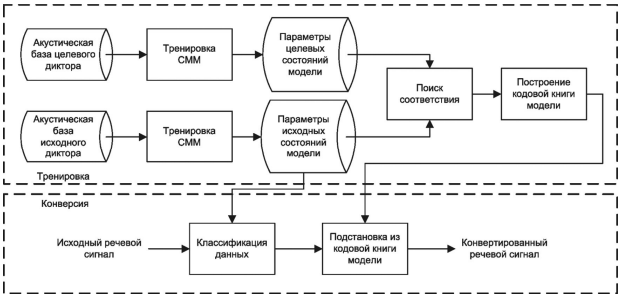


Рис. 1. Общая структура системы конверсии с текстонезависимым обучением

Две различные скрытые марковские модели отдельно тренируются на исходной и целевой акустической базах. Модели описывают основные статистические характеристики голосов дикторов. Преобразование данных характеристик выполняется путем нахождения соответствия между исходной и целевой моделями. Традиционно функция конверсии ищется в виде линейного преобразования, однако это приводит к проблеме чрезмерного сглаживания. Для решения данной проблемы в [6] предлагается метод подстановки. При использовании данного метода множество состояний каждой СММ используется в качестве одной заменяемой структурной единицы. После тренировки СММ для исходного и целевого дикторов, выполняется поиск соответствия между структурами (паттернами) состояний, которые описываются при помощи распределения Гаусса.

При выполнении конверсии каждый фрейм заданной входной последовательности помечается индексом состояния на основании распределений вероятности состояний исходной СММ. Таким образом, формируется последовательность входных индексов, которая отражает последовательность характеристических векторов, учитывающую статистические параметры модели. После чего выполняется подстановка, в результате которой каждый исходный индекс заменяется на соответствующий ему целевой в соответствии с кодовой книгой. Для того, чтобы выходная последовательность векторов была более естественной используется специальный синтез на основе СММ. Важной особенностью текстонезависимого подхода конверсии является большая (по сравнению с тектозависимым подходом) речевая база, необходимая для обучения. В частотности, в работе [6] для обучения использовалось 180 предложений для каждого из дикторов. Как правило, конверсия выполняется не в реальном масштабе времени, позволяющем применять фонетический анализ речи не только для обучающих речевых баз, но и для самого конвертируемого речевого сигнала [3], [5-6].

Предположим, что выполняется автоматическая сегментация обучающих речевых баз на фонетические единицы (фонемы и/или аллофоны) с последующим поиском соответствий в исходной и целевой последовательностях, и что можно поставить характеристические векторы целевого и исходного дикторов во взаимооднозначное соответствие при помощи кодовой книги. В таком случае, становится возможной независимая конверсия отдельных фреймов речевого сигнала и, следовательно, обработка речи в реальном масштабе времени.

Для реализации вышеизложенной идеи в настоящей работе предлагается текстонезависимая система конверсии голоса, структура которой представлена на рис. 2. Система включает следующие функциональные блоки: блок параметрического описания речевого сигнала (гармонический или синусоидальный анализатор, средство параметрического моделирования шума), блок сегментации речевых баз на фонетические единицы, блок кластеризации и построения кодовых книг, блок конверсии параметров и блок синтеза выходного речевого сигнала.

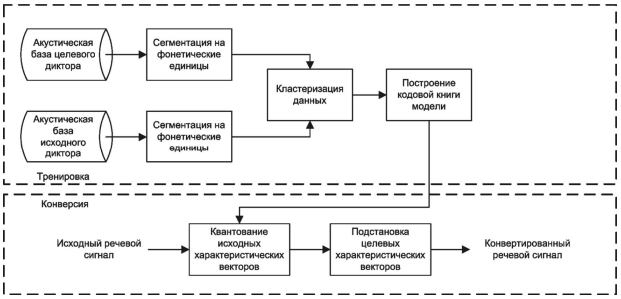


Рис. 2. Общая структура предлагаемой системы конверсии с текстонезависимым обучением

Сегментация речевых баз и вся последующая обработка (в том числе и сама конверсия голоса) выполняются с использованием параметрического описания сигнала. Наиболее подходящей моделью представления в задачах синтеза и конверсии голоса традиционно считается гибридная модель, обеспечивающая раздельное описание периодической (квазистационарной или детерминистской) и непериодической (нестационарной или стохастической) составляющих [8-9]. Параметрическое описание выполняется путем оценки кепстральных коэффициентов либо оценки линейных спектральных частот LSF (line spectral frequencies), причем, в силу своих свойств, последний способ представления предпочтительнее при использовании в процессе конверсии техники векторного квантования. Значения LSF вычисляются из коэффициентов линейного предсказания, которые, как правило, оцениваются на фрейме анализа автокор-

реляционным методом. Данный способ оценки имеет существенные ограничения, обусловленные свойствами линейного предсказания. С одной стороны, для того чтобы избежать значительного усреднения (сглаживания) спектральной огибающей следует использовать большое число коэффициентов предсказания, с другой стороны увеличение числа коэффициентов приводит к потере ее формы и моделированию отдельных гармоник.

Для повышения точности параметрического описания вокализованной речи в работе используется оригинальный способ оценки коэффициентов предсказателя, основанный на преобразовании мгновенных гармонических параметров. Это позволяет получать локализованные по времени оценки спектральных огибающих и использовать для этой цели предсказатели высоких порядков.

Параметризация речевого сигнала

В системах конверсии голоса для параметрического описания речи существует две основные альтернативы: линейное предсказание и синусоидальное моделирование (как основной способ, либо как часть гибридного описания). Гибридные модели (такие как гармоники+шум либо синусоиды+шум) являются более современными, и большинство исследователей сходятся во мнении, что они так же имеют ряд преимуществ перед моделями на основе линейного предсказания. С другой стороны, чистое синусоидальное представление речевого сигнала лишено физической интерпретации в том смысле, что оно напрямую не связано с параметрами речевого тракта, в отличие от линейного предсказания. Поскольку для преобразования голоса необходимо рассматривать спектральные огибающие как отдельные самостоятельные единицы и в то же время сохранить возможность параметрического синтеза отдельных гармоник, в работе предлагается использовать синусоидальное описание совместно с линейным предсказанием. Разделение речевого сигнала на периодическую и непериодическую составляющие выполняется при помощи гармонического анализа и синтеза, после чего каждая из составляющих описывается при помощи LSF. Причем, для оценки LSF периодической составляющей используются мгновенные гармонические параметры, а для оценки LSF шумовой – традиционный автокорреляционный метод.

Так как синусоидальное моделирование в большинстве случаев можно применить непосредственно к входу системы, входной сигнал удобно рассматривать как комбинацию периодической и остаточной компонент. Таким образом, сигнал $s(n)$ можно записать в виде соотношения:

$$s(n) = \sum_{k=1}^{K} \mathrm{MAG}_k(n) \cos \varphi_k(n) + r(n) \qquad (1)$$

где $\mathrm{MAG}_k(n)$ – мгновенная амплитуда k-й синусоиды; K – число синусоид; $\varphi_k(n)$ – мгновенная фаза k-й синусоиды; $r(n)$ – сигнал-остаток.

Мгновенная фаза $\varphi_k(n)$ и мгновенная частота $f_k(n)$ соотносятся следующим образом:

$$\varphi_k(n) = \sum_{i=0}^{n} \frac{2\pi f_k(i)}{F_s} + \varphi_k(0) \qquad (2)$$

где F_s – частота дискретизации; $\varphi_k(0)$ – начальная фаза k-й синусоиды.

Наряду с синусоидальной моделью широко применяется (особенно в вокодерных системах) гармоническая модель, которая предполагает, что значения мгновенных частот $f_k(n)$ являются кратными частоте основного тона $f_0(n)$ и могут быть вычислены по следующей формуле:

$$f_k(n) = k f_0(n) \qquad (3)$$

Гармоническая модель используется для кодирования речевых сигналов с высоким коэффициентом сжатия, так как обеспечивает чрезвычайно эффективное описание вокализованных фрагментов речи.

Считается, что амплитуда и фаза синусоидальных компонент изменяются медленно, поэтому можно сделать следующие предположения:

- каждая синусоида может быть ограничена в частотной области узкой частотной полосой;
- синусоидальные компоненты разделены в частотной области (их можно выделить на всем протяжении анализируемого фрейма фильтрами с неперекрывающимися полосами пропускания), в противном случае они создают переходную (транзиентную) компоненту;
- синусоидальные компоненты достаточно продолжительны, в противном случае они формируют либо переходную компоненту, либо шум.

Таким образом, искомые параметры синусоидальной модели $\mathrm{MAG}_k(n)$ и $f_k(n)$ являются гладкими, непрерывными функциями с ограниченным частотным диапазоном. Разделение сигнала на периодическую и остаточную части, как и оценка гармонических параметров, является фундаментальной задачей синусоидального моделирования. Точность оценок, как правило, оказывает существенное влияние на качество работы систем, что свидетельствует о необходимости совершенствования методов гармонического анализа. Неточное разделение вносит в обрабатываемый сигнал слышимые артефакты, которые затем, на последующих стадиях обработки, не могут быть исключены. Для разделения сигнала на периодическую и шумовую составляющие в данной работе используется метод, выполняющий оценку мгновенных гармонических параметров [8].

Фильтр анализа

Для оценки гармонических параметров речи используется частотно-модулированный фильтр анализа. Узкополосная фильтрация, с одной стороны, обеспечивает разделение анализируемого сигнала на периодические компоненты и получение мгновенных гармонических параметров, с другой – позволяет правильно обрабатывать компоненты с частотной модуляцией. Данный подход объединяет в себе элементы преобразований со встроенным временным масштабированием [10-11] и методов оценки на основе аналитических сигналов.

Рассматривая центральную частоту полосы пропускания, как функцию от дискретного времени $F_c(n)$, вы-

ражение импульсной характеристики фильтра анализа может быть представлено в следующем виде [19]:

$$h(n)=\begin{cases}1, & n=0\\ \dfrac{F_s}{n\pi}\cos\left(\dfrac{2\pi}{F_s}\varphi_c(n,i)\right)\sin\left(\dfrac{2\pi n}{F_s}F_\Delta\right) & n\neq 0\end{cases}\Big/\,2F_\Delta \qquad (4)$$

где $2F_\Delta$ – ширина полосы пропускания фильтра и

$$\phi_c(n,i)=\begin{cases}\sum_{j=n}^{i}F_c(j), & n<i;\\ -\sum_{j=i}^{n}F_c(j), & n>i;\\ 0, & n=i.\end{cases}$$

Выходной сигнал фильтра $S_{F_cF_\Delta}$, который является сверткой входного сигнала с импульсной характеристикой, представляет собой периодический сигнал, параметры которого (мгновенная амплитуда $\mathrm{MAG}(n)$, фаза $\varphi(n)$ и частота $f(n)$ могут быть вычислены при помощи следующих формул (индекс n соответствует номеру отсчета входного сигнала, а i – номеру отсчета импульсной характеристики):

$$S_{F_cF_\Delta}=\mathrm{MAG}(n)\cos(\varphi(n)) \qquad (5)$$

$$\mathrm{MAG}(n)=\sqrt{A^2(n)+B^2(n)}\,; \qquad (6)$$

$$\varphi(n)=\arctan\left(\frac{-B(n)}{A(n)}\right); \qquad (7)$$

$$f(n)=\frac{\varphi(n+1)-\varphi(n)}{2\pi}F_s\,. \qquad (8)$$

где $A(n)=\sum_{i=0}^{N-1}\dfrac{s(i)F_s}{2\pi(n-1)F_\Delta}\sin\left(\dfrac{2\pi(n-i)}{F_s}F_\Delta\right)\cos\left(\dfrac{2\pi}{F_s}\varphi_c(n,i)\right)$;

$B(n)=\sum_{i=0}^{N-1}\dfrac{-s(i)F_s}{2\pi(n-1)F_\Delta}\sin\left(\dfrac{2\pi(n-i)}{F_s}F_\Delta\right)\cos\left(\dfrac{2\pi}{F_s}\varphi_c(n,i)\right)$.

Частотно-модулированный (ЧМ) фильтр имеет масштабированную в частотной области полосу пропускания, задаваемую частотным контуром $F_c(n)$, что обеспечивает анализ периодических компонент с быстрым изменением частоты в узкой полосе – рис. 3.

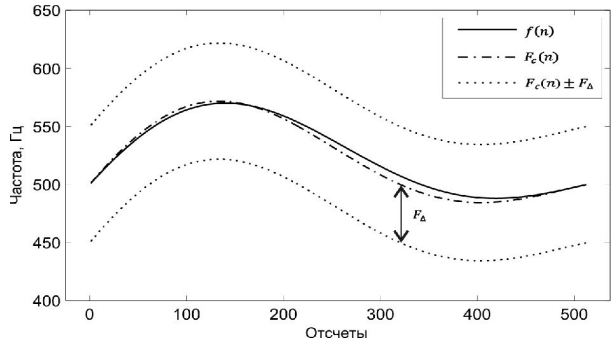


Рис.3. Кратковременный гармонический анализ с использованием частотно-модулированного фильтра

Сужение полосы делает оценки параметров более точными и позволяет применять мгновенный гармонический анализ к гармоникам высокого порядка - чем выше номер гармоники, тем больше изменение ее частоты, и импульсная характеристика ЧМ фильтра изменяется соответствующим образом.

Оценка мгновенных гармонических параметров речи

Учитывая специфику задачи конверсии голоса, требуется компактное описание речевого сигнала, наиболее удачно комбинирующее возможности описания звуков различной природы (вокализованных, взрывных и невокализованных). Для кодирования и обработки звуковых сигналов, как было сказано выше, используются гибридные модели, разделяющие сигнал на периодическую, транзиентную и шумовую компоненты. Первичное разделение синусоидальной (периодической) части и остатка может выполняться путем гармонического анализа на основе фильтров.

Вначале синусоидальный анализатор выполняет оценку параметров с постобработкой для выделения продолжительных периодических компонент в низкочастотной полосе сигнала. Локализация периодических компонент в частотной области выполняется путем итеративного перерасчета. На каждом шаге полоса пропускания фильтра перемещается в соответствии с полученными оценками для того, чтобы поместить локальный максимум энергии оцениваемой компоненты в центре полосы [8]. Короткие и нестабильные спектральные компоненты отбрасываются путем слежения за полученными синусоидальными параметрами. Для выделения достаточно продолжительных и стабильных значений сравниваются частоты и амплитуды компонент соседних фреймов.

Из выделенных спектральных компонент выбирается одна, которая соответствует основному тону речи. Частотный контур основного тона определяется так, чтобы найти приблизительные частотные траектории гармоник, необходимые для того, чтобы синтезировать ЧМ фильтры анализа.

Необходимые траектории центральных частот полос пропускания фильтров $F_c(n)$ вычисляются как мгновенная частота основного тона, умноженная на номер k соответствующей гармоники $F_c^k(n)=kf_0(n)$. Процедура оценки проводится последовательно, начиная с первой гармоники и заканчивая последней. После оценки каждой следующей гармоники контур частоты основного тона уточняется с учетом полученных параметров по следующей формуле перерасчета:

$$f_0(n)=\sum_{i=0}^{k}\frac{f_i(n)\mathrm{MAG}_i(n)}{(i+1)\sum_{j=0}^{k}\mathrm{MAG}_j(n)}\,. \qquad (9)$$

Таким образом, при оценке гармоник высокого порядка значения частоты основного тона становятся более точными, что позволяет правильно синтезировать частотно-модулированный фильтр анализа.

Оценка линейных спектральных частот из мгновенных гармонических параметров речи

Существует большое количество работ, посвященных способам вычисления параметров предсказания, получения с их помощью спектральных оценок и определения формантных траекторий [12-15]. Параметры

модели линейного предсказания (далее коэффициенты предсказания) могут быть описаны через коэффициенты отражения секций акустической трубы, что обеспечивает возможность определения параметров голосового тракта диктора. Это свойство может применяться для диагностики различных заболеваний.

Модель линейного предсказания основывается на предположении, что любой отсчет речевого сигнала $s(n)$ можно приближенно оценить линейной комбинацией некоторого числа p предшествующих ему отсчетов, что приводит к следующему соотношению:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \cdot \quad (10)$$

где a_1, a_2, ..., a_p – коэффициенты предсказания; $u(n)$ – нормализованная последовательность возбуждения (ошибка предсказания); G – коэффициент усиления [15].

В z-области коэффициенты предсказания задают передаточную функцию

$$H(z) = \frac{G}{1 - \sum_i^p a_i z^{-i}} = \frac{G}{A(z)} \cdot \quad (11)$$

Ошибка предсказания $e(n)$ определяется как разность между исходными и приближенно вычисленными (предсказанными) отсчетами:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \cdot \quad (12)$$

Основная задача ЛП – определение набора коэффициентов предсказания, которые минимизируют $e(n)$.

Существуют два основных метода определения коэффициентов предсказания: автокорреляционный и ковариационный. Оба метода используют представление сигнала во временной области.

Автокорреляционный метод предполагает, что $s(n) = 0$ вне заданного сегмента сигнала $0 \le n < N$, минимизирует ошибку предсказания на бесконечном интервале и сводится к решению системы:

$$\sum_{i=1}^{p} a_i r(|i-j|) = -r(i) \cdot \quad (13)$$

где $j = 1, 2, ..., p$; $r(l) = \sum_{n=0}^{N-1-l} s(n)s(n+1)$ – автокорреляционная функция, $l \ge 0$.

Для оценки коэффициентов линейного предсказания из мгновенных гармонических параметров используется метод, изложенный в [15]. Целевая спектральная огибающая рассматривается как непрерывная функция амплитуды от частоты $\mathrm{MAG}(\omega)$, заданная на интервале [0, π]. Вычисление элементов матриц $r(l)$ системы (13) выполняется при помощи следующего выражения:

$$r(l) = \int_0^{\pi} MAG(\omega)\cos(\omega l)d\omega. \quad (14)$$

Если функция $\mathrm{MAG}(\omega)$ содержит точки разрыва $\omega_d = (\omega_1, \omega_2, ..., \omega_l)$, тогда выражение принимает следующую форму:

$$r(l) = \sum_{i=1}^{I+1} \int_{\bar{\omega}d,i}^{\bar{\omega}d,i+1} \mathrm{MAG}(\omega)\cos(\omega l)d\omega \quad (15)$$

где $\bar{\omega}_d = (0, \omega_1, \omega_2, ..., \omega_l, \pi)$.

Непрерывная амплитудная огибающая спектра может быть получена из векторов амплитудных и частотных значений путем линейной интерполяции. Каждый сегмент огибающей $f_i \le \omega \le f_{i+1}$, $1 \le i \le K-1$ описывается линейным уравнением прямой $\mathrm{MAG}(\omega) = b_i\omega + c_i$. Параметры b_i и c_i вычисляются из смежных значений амплитуды и частоты. Элементы системы автокорреляционного метода (13), определяемые соотношением (15), вычисляются следующим образом:

$r(l) = \sum_{i=1}^{K-1} D(l,i)$, где

$$D(l,i) = \begin{cases} \frac{b_i}{l^2}[\cos(f_{i+1}l) + f_{i+1}l\sin(f_{i+1}l)] + \frac{c_i}{l}\sin(f_{i+1}l) - \\ -\frac{b_i}{l^2}[\cos(f_i l) + f_i l\sin(f_i l)] - \frac{c_i}{l}\sin(f_i l), \text{ если } l \ne 0; \\ \frac{1}{2}b_i f_{i+1}^2 + c_i f_{i+1} - \frac{1}{2}b_i f_i^2 + c_i f_i, \text{ если } l = 0. \end{cases}$$

Сегментация речевой базы и формирование кодовой книги функции конверсии

Сегментация обучающих последовательностей

Основным инструментом автоматической сегментации речи является алгоритм K-средних на основе скрытой марковской модели. СММ представляет собой набор состояний, связанных между собой переходами. Каждому переходу соответствует определенная вероятность a_{ij}, а каждому состоянию q_n – вероятность $b_i(u)$ появления символов $O_1, O_2, ..., O_M$, в данном состоянии [15].

Допустим, что имеется тренировочная последовательность наблюдений (параметрическая речевая база в виде характеристических векторов) и первоначальная оценка параметров модели. Данное первое приближение может быть выбрано случайным образом, либо на основе доступной информации об имеющихся данных. После инициализации модели множество тренировочных наблюдений разбивается на состояния в соответствии с текущими параметрами модели λ. Разбиение выполняется при помощи алгоритма Витерби и алгоритма обратного хода по оптимальному пути.

Результатом сегментации каждой из тренировочных фраз является нахождение для каждого из N состояний наиболее вероятной (для текущих параметров модели) последовательности наблюдений, которая присутствует внутри каждого состояния j. Исходя из текущего разделения на состояния, рассчитываются новые значения коэффициентов a_{ij} путем подсчета числа переходов из состояния i в состояние j и деления его на общее число переходов из состояния i (включая переходы в само состояние i). Из полученных параметров получается новая СММ $\hat{\lambda}$ при помощи стандартной процедуры перерасчета.

Для выполнения сегментации следует определить число состояний, равное числу различаемых моделью фонетических единиц.

Преобразование букв в фонемы

При сегментации тренировочных речевых предложений используется подстрочный текст, который должен быть трансформирован в набор фонем русского языка для того, чтобы соответствовать фонетическим состояниям, различаемым моделью. В работе использована методика перевода букв в фонемы, изложенная в [17]. В экспериментальной реализации системы конверсии использовался набор из 30-и фраз со следующим фонетическим содержанием – рис.4.

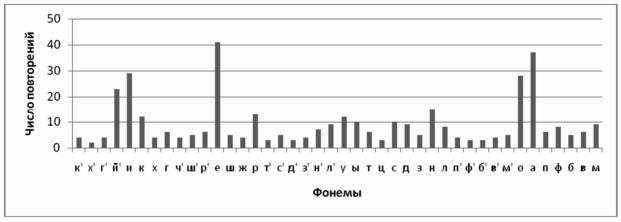


Рис. 4. Фонетическое содержание набора обучающих фраз

Поиск функции конверсии при помощи векторного квантования

В задачах конверсии голоса техника векторного квантования определяется специфическими критериями (отличными от тех, которые применяются при кодировании речи) и выбирается исходя из заданных условий. Важным вопросом так же является выбор параметров для квантования и выбор числа уровней.

Исходя из задачи конверсии голоса в режиме реального времени, необходимым свойством является возможность определения индекса вектора кодовой книги целевого диктора, соответствующего определенному вектору кодовой книги исходного диктора. Учитывая, что в процессе обучения выполняется сегментация речевой базы на фонетические единицы, наиболее очевидным решением является построение кодовой книги таким образом, чтобы ее кластеры были фонетически мотивированы, т.е. обеспечивали наименьшее расстояние между центроидом кластера и векторами, принадлежащими соответствующей фонеме.

В случае, когда число уровней кодовой книги равно числу фонем, кодовая книга может быть получена непосредственно путем вычисления центроидов фонем. Учитывая, что доступная обучающая выборка ограничена и не содержит большого числа повторений одной и той же фонемы, данный способ построения кодовой книги может оказаться достаточно эффективным.

Если число уровней кодовой книги превышает число фонем, то целесообразно разбить фонемные кластеры на подкластеры, используя какую-либо дополнительную информацию (например энергию или значения частоты основного тона).

В ходе данной работы были реализованы системы конверсии с размером кодовой книги 42, 84, 126, 256 и 512 векторов (использовались векторы LSF с размерностью 30), применяя различные способы кластеризации (разбиение на акустические классы, максимизация вероятности соответствия и др.). Был сделан вывод, что на заданной обучающей выборке (30 фраз) наиболее предпочтительным является использовать маленькую кодовую книгу (42 вектора). Кодовые книги с большим числом уровней позволяют получить меньшую ошибку квантования и, соответственно, могут обеспечить более высокое качество реконструированного сигнала. Однако данное преимущество не может быть в полной мере использовано из-за ошибки сегментации, которая в той или иной мере обязательно присутствует в процессе обучения. В случае, когда большинство фонем обучающей выборки повторяются 3-10 раз невозможно набрать достаточное число обучающих векторов для того, чтобы при помощи статистических методов избавиться от этой ошибки. Был сделан вывод, что для использования большей кодовой книги в системе текстонезависимой конверсии голоса требуется существенно более продолжительная обучающая последовательность.

Результаты экспериментов

Общая структура экспериментальной системы конверсии голоса

Система конверсии голоса может быть разделена на три основных функциональных блока: блок анализа, блок обработки и блок синтеза – рис. 5. Блок гармонического анализа выполняет оценку гармонических параметров с разделением входного сигнала на периодическую и остаточную части. Все модификации голоса выполняются в блоке обработки, причем периодическая и остаточная части обрабатываются отдельно. Выходной сигнал системы представляет собой сумму модифицированных периодической и шумовой компонент, которые синтезируются из полученных целевых параметров.

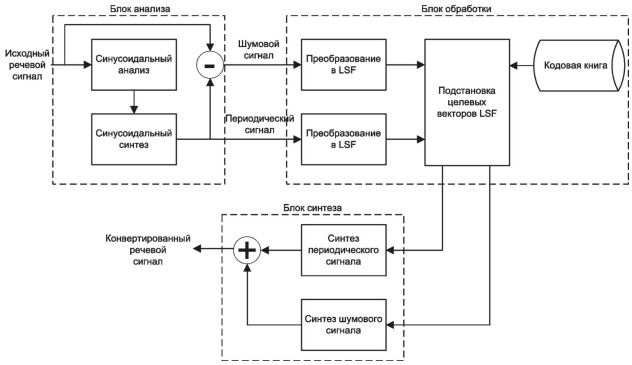


Рис. 5. Система конверсии голоса

Как было сказано выше, конверсия голоса осуществляется в два этапа: обучение, т.е. формирование правил конверсии параметров речи исходного диктора в параметры целевого диктора, и непосредственно конверсия голоса путем применения сформированных во время обучения правил.

Индивидуальность диктора составляют два основных фактора: акустические и просодические свойства речи [18]. Для осуществления качественной конверсии голоса необходимо контролировать и те и другие свойства, что достигается путем использования синусоидальной модели речи, так как она обеспечивает контроль и эффективное изменение тембра, частоты основного тона, длительности фонетических единиц и общего темпа речи.

Обработка спектральных огибающих

Спектральные огибающие являются важной характеристикой голоса, так как они содержат информацию о тембре и индивидуальности диктора. Голосовая речь может быть представлена в виде набора таких огибающих и контура частоты основного тона. Поэтому оценка и модификация спектральных огибающих позволяет создавать специальные эффекты, такие как конверсия голоса и изменение тембральной окраски. Гармоническая модель, описывающая сигнал в виде суммы синусоид различной амплитуды и частоты, позволяет оценить мгновенную спектральную огибающую при помощи интерполяции [19].

Исходный вокализованный фрагмент анализируется гармоническим анализатором, затем огибающие спектра вычисляются с использованием мгновенных значений гармонических амплитуд и частоты основного тона.

Спектральные огибающие переводятся в LSF, после чего выполняется поиск вектора в кодовой книге исходного диктора наиболее близкого к полученному. Затем из кодовой книги целевого диктора выбирается LSF вектор с соответствующим индексом, который определяет целевую спектральную огибающую.

Изменение основного тона

Спектральные огибающие характеризуют тембр звучания голоса, в то время как основной тон характеризует интонацию. Основной тон, как и тембр, является характеристикой только вокализованных звуков. Соответственно модификация основного тона выполняется только для гармонической части сигнала. В задачах конверсии голоса, считают, что основной тон может изменяться на основе статистических закономерностей.

В реализованной системе требуемый новый контур основного тона определяется при помощи алгоритма нормализации Гаусса. Метод основан на приведение в соответствие математического ожидания значений частоты основного тона и среднего отклонения исходного и целевого дикторов [18]. Конвертированное значение частоты основного тона $p_t^{S\to T}$ вычисляется как

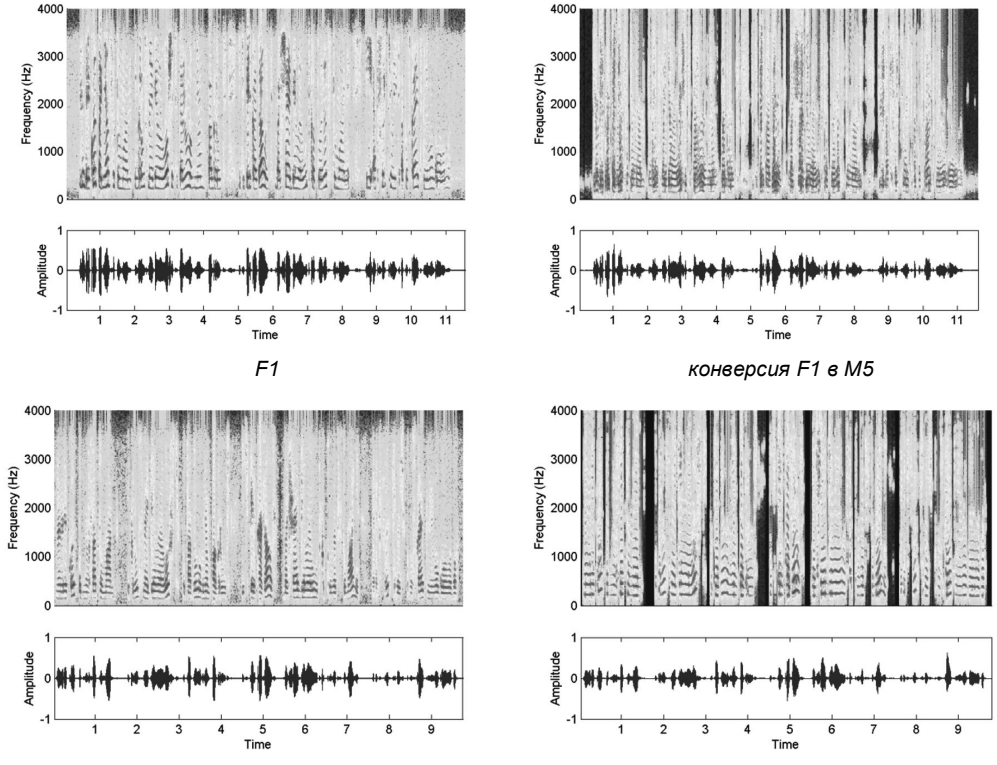
$$p_t^{S\to T} = \frac{p_t^S - \mu^s}{\sigma^s}\sigma^T + \mu^T \qquad (17)$$

где μ^s и σ^s – математическое ожидание и среднее отклонение частоты основного тона исходного диктора соответственно; μ^T и σ^T – математическое ожидание и среднее отклонение частоты основного тона целевого диктора соответственно; p_t^S – заданная частота основного тона исходного диктора.

Параметры μ^s, σ^s, μ^T и σ^T хранятся в кодовой книге.

Результаты экспериментов

Для оценки качества описанных методов конверсии голоса на основе специальной речевой базы данных выполнена экспериментальная конверсия голоса. База данных содержит записи речи 14 различных дикторов (7 мужских и 7 женских голосов, условно обозначенных M1-7 и F1-7 соответственно) с частотой дискретизации 8 кГц. Для обучения кодовых книг использовались специальные фразы, обеспечивающие фонетическую сбалансированность выборки. Результат конверсии получен для каждой пары исходный-целевой диктор. Пример полученных образцов речи приведен ниже – рис. 6. Для того, чтобы оценить вклад предложенного способа обучения, речевая база была обработана системой с текстозависимым обучением [20].



F1 *конверсия F1 в M5*

M5 *конверсия M5 в F1*

Рис. 6. Результат конверсии голоса

Субъективная оценка качества конверсии выполнялась путем прослушивания полученных звуковых файлов группой экспертов из 5-и человек. Каждому эксперту было предложено оценить разборчивость и узнаваемость по шкале от 0% (полное отсутствие) до 100% (абсолютная натуральность). Оценки были сгруппированы по направлению конверсии (мужской в мужской, мужской в женский, женский в мужской, женский в женский - условно группы обозначены MM, MF, FM, FF соответственно) и внутри каждой группы по типу обучения (текстозависимое/текстонезависимое). По каждой из подгрупп вычислены средние значения оценок – табл. 1.

Таблица 1

Средние показатели качества конверсии голоса

	Конверсия с текстозависимым обучением		Конверсия с текстонезависимым обучением	
Направление конверсии	Разборчивость %	Узнаваемость %	Разборчивость %	Узнаваемость %
MM	62	68	58	65
MF	66	72	70	70
FM	80	84	74	82
FF	84	80	78	82
Среднее	73	76	70	75

Эксперимент показал, что полученное качество конверсии, учитывая уровень слышимых артефактов и натуральность звучания близко к результатам, получаемым при использовании системы конверсии с текстозависимым обучением. Средний показатель разборчивости составляет 70%, а средний показатель узнаваемости – 75%.

Заключение

В работе предложена схема конверсии голоса с текстонезависимым обучением, использующая специальную модель для параметрического описания речевого сигнала. Особенность модели заключается в способе оценки мгновенных значений LSF высокого порядка, определяющих форму спектральных огибающих вокализованных фрагментов речи. Оценка производится в два этапа: сначала выполняется оценка мгновенных гармонических параметров, затем выполняется их преобразование в параметры модели линейного предсказания.

Конверсия выполняется раздельно для спектральных огибающих и контура основного тона речевого сигнала. В процессе конверсии спектральные огибающие заменяются на целевые, для чего используется кодовая книга конверсии. Основной тон модифицируется при помощи алгоритма нормализации Гаусса. Кодовая книга конверсии формируется в процессе обучения системы. Для обучения используются образцы исходного и целевого дикторов, которые сегментируются и сопоставляются при помощи скрытой марковской модели.

Результаты практического применения свидетельствуют о том, что используя предлагаемые методы, возможно создание текстонезависимой системы конверсии голоса, причем качество реконструкции сигнала может быть близким к тому, которое обеспечивают текстозависимые системы конверсии. Для заметного повышения показателей разборчивости и узнаваемости конвертированной речи следует вести исследования в направлении расширения фонетического базиса со значительным увеличением обучающей выборки.

Литература

1. Dutoit, T. An Introduction to Text-to-speech Synthesis / T. Dutoit. - The Netherlands: Kluwer Academic Publishers, 1997. – 326 p.

2. Abe, M. Voice conversion through vector quantization / M. Abe, S. Nakamura, K Shikano. // Acoustics, Speech, and Signal Processing: proceedings of int. conf. (ICASSP-88), New York, USA, April 1-14, 1988. - New York, 1988. - P. 655-658.

3. Erro, D. On combining statistical methods and frequency warping for high-quality voice conversion / D. Erro, T. Polyakova and A. Moreno // Acoustics, Speech, and Signal Processing: proceedings of int. conf. (ICASSP-2008 Las Vegas, USA, March 30- April 4, 2008. – Las Vegas, 2008. - P. 4665-4668.

4. Azarov, E. Text and speaker independent voice conversion / E. Azarov, A. Petrovsky // Pattern recognition and information processing: proceedings of the 10-th intern. conf., Belarus, Minsk, May 19–21, 2009. – Minsk, 2009. – P. 195–198.

5. Sundermann, D. Text-independent voice conversion based on unit selection/ D. Sundermann, [et al.] // Acoustics, Speech, and Signal Processing: proceedings of int. conf. (ICASSP-2006), Toulouse, France, May 15-19, 2006. – Toulouse, 2006. - P. 81-84.

6. Zhang, M. Text-independent voice conversion based on state mapped codebook / M. Zhang, [et al.] // Acoustics, Speech, and Signal Processing: proceedings of int. conf. (ICASSP-2008), Las Vegas, USA, March 30- April 4, 2008. – Las Vegas, 2008. - P. 4605-4608.

7. Erro, D., INCA algorithm for training voice conversion systems from nonparallel corpora / D. Erro, A. Moreno and A. Bonafonte // IEEE transactions on audio, speech, and language processing – 2010. - Vol.18, № 5. – P. 944-953.

8. Petrovsky Al., Azarov E. and Petrovsky A., Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding // Signal Processing, Volume 91, Issue 6, Fourier Related Transforms for Non-Stationary Signals, pp. 1489-1504, June 2011.

9. Азаров, И.С. Мгновенный гармонический анализ: обработка звуковых и речевых сигналов в системах мультимедиа / И.С. Азаров, А.А. Петровский - LAP Lambert Academic Publishing, Saarbrucken, 2011. – 163 с.

10. Weruaga, L. The fan-chirp transform for non-stationary harmonic signals / L. Weruaga, M. Kepesi // Signal Processing. – 2007. - Vol.87, № 6. – P. 1-18.

11. Zhang, F. Harmonic transform / F. Zhang, G. Bi, Y.Q. Chen // IEEE Proc.-Vis. Image Signal Process. – 2004. - Vol.151, № 4. – P. 257-264.

12. Huang, X. Spoken language processing / X . Huang, A. Acero, H.W. Hon. - New Jersey: Prentice Hall, 2001. – 1008 p.

13. Kondoz, A.M. Digital speech: coding for low bit rate communication systems / A.M. Kondoz – New York: John Wiley & Sons Inc., 1996. – 442 p.

14. Оппенгейм, А. Цифровая обработка сигналов / А. Оппенгейм, Р. Шафер. – Техносфера, 2006. – 858 с.

15. Rabiner, L.R. Fundamentals of speech recognition / L.R. Rabiner, B.H. Juang. - New Jersey: Prentice Hall, 1993. - 496 p.

16. Azarov, E., Petrovsky, A. «Linear prediction of deterministic components in hybrid signal representation», Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS), pp.2662-2665, Paris May 30 2010-June 2 2010.

17. Лобанов, Б.М. Компьютерный синтез и клонирование речи / Б.М. Лобанов, Л.И. Цирульник. – Минск «Белорусская наука» 2008. – 344 с.

18. Lee, K., Statistical conversion algorithms of pitch contours based on prosodic phrases / K. Lee, Y Zhao // Speech Prosody 2004: proceedings of the int. conf. (SP 2004), Nara, Japan, March 23-26, 2004, CD-ROM.

19. Instantaneous harmonic analysis for vocal processing [electronic resource] / E. Azarov, A. Petrovsky. - DAFx-09: proc. of the 12th International Conference on Digital Audio Effects, Italy, Como, September 1-4, 2009. – Como., 2009. – Mode of access: http://dafx09.como.polimi.it /proceedings/papers/paper_25.pdf. - Date of access: 04.09.2009.

20. Azarov, E.; Petrovsky. A. «Real-time voice conversion based on instantaneous harmonic parameters» Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2011), pp. 5140 - 5143, Prague, Czech Republic, May 22-27, 2011.

REAL-TIME VOICE CONVERSION SYSTEM WITH TEXT-INDEPENDENT TRAINING BASED ON HYBRID PARAMETRIC REPRESENTATION OF SPEECH

Azarov E., Petrovsky A.

The paper presents a voice conversion technique based on hybrid parametric speech representation. The conversion function is estimated from speech samples of the source and target speakers. The samples have arbitrary text content. Effectiveness of the proposed technique is rated by experimental implementation of the voice conversion system.

14-я Международная научно-техническая конференция и выставка **«Цифровая обработка сигналов и её применение - DSPA'2012»**

Обработка сигналов в радиотехнических и информационно-измерительных системах

Секция 4: Обработка сигналов в радиотехнических системах (Руководители – д.т.н., профессор Ю.Г. Сосулин, к.т.н., профессор В.С. Сперанский)

Обсуждение проблем и задач обработки радиотехнических сигналов велось по трем основным направлениям: обнаружение сигналов и оценивание их параметров; радиолокация, пеленгация и навигация; пространственно-временная обработка сигналов.

По итогам прошедшей конференции были представлены к награждению Дипломами лауреатов конкурса молодых ученых следующие работы:

1. Применение методов Прони и Штейглица-Макбрайда для формирования весовых коэффициентов при адаптивной фильтрации неклассифицированных выборок наблюдения. Автор: Гордеев А.Ю., аспирант ОАО «ВНИИРТ».

2. Статистический синтез и сравнительный анализ оценок корреляционной размерности. Автор: Паршин А.Ю., студент Рязанского государственного радиотехнического университета.

3. Аналого-цифровое преобразование в цифровых антенных решетках. Автор: Бохин Д.Л., аспирант Московского авиационного института