УДК 004.934

И.С. Азаров, А.А. Петровский

ФОРМИРОВАНИЕ ПЕРСОНАЛЬНОЙ МОДЕЛИ ГОЛОСА ДИКТОРА С УНИВЕРСАЛЬНЫМ ФОНЕТИЧЕСКИМ ПРОСТРАНСТВОМ ПРИЗНАКОВ НА ОСНОВЕ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТИ

Азаров И.С., Петровский А.А. **Формирование персональной модели голоса диктора с универсальным фонетическим пространством признаков на основе искусственной нейронной сети.**

Аннотация. В работе исследуется возможность формирования модели голоса заданного диктора на основе записей образцов его голоса с транскрипцией. В работе предлагается практический способ построения голосовой модели и результаты экспериментов ее применения к задаче конверсии голоса. Модель использует искусственную нейронную сеть, устроенную по принципу автоматического кодера, устанавливающую соответствие между пространством речевых параметров и пространством возможных фонетических состояний, унифицированным для произвольного голоса.

Ключевые слова: конверсия голоса, синтез речевого сигнала, искусственная нейронная сеть.

Azarov E., Petrovsky A. **Training Personal Voice Model of a Speaker with Unified Phonetic Space of Features Using Artificial Neural Network.**

Abstract. The paper investigates possibility of creating a personal voice model using transcribed speech samples of a specified speaker. The paper presents a practical way of building such speech model and some experimental results of applying the model to voice conversion. The model uses an artificial neural network organized as autoencoder that establishes correspondence between space of speech parameters and space of possible phonetic states, unified for any voice.

Keywords: voice conversion, speech synthesis, artificial neural network.

1. Введение. В настоящее время большую часть существующих задач обработки речевых сигналов можно условно разделить на следующие основные направления: синтез речи по тексту, распознавание речи, создание различных звуковых эффектов (например, конверсия голоса), кодирование речи и улучшение речевых характеристик (например, повышение разборчивости речи и шумоподавление).

Синтез речи по тексту и распознавание речи являются, вероятно, наиболее важными из всех, поскольку в перспективе могут привести к организации полноценного голосового интерфейса между человеком и компьютером. Последние коммерческие решения в данных областях являются многообещающими: представлены синтезаторы речи по тексту (http://www.speechpro.ru/product/recognition/tts, https://play.google.com/store/apps/details?id=com.google.android.tts&hl=ru), позволяющие генерировать практически натуральную речь с низким уровнем слышимых артефактов, и современные распознаватели, по-

зволяющие распознавать фразы слитной речи с недостижимой ранее высокой вероятностью. Основной причиной такого заметного прогресса является появление возможности использования огромных речевых корпусов, обусловленное многократным удешевлением и ускорением вычислительного процесса. Так, для качественного синтеза речи по тексту теперь вместо аллофонного синтеза, использующего речевую базу аллофонов общей продолжительностью в несколько минут, применяется корпусный синтез, использующий десятки часов речевого материала состоящий из всевозможных комбинации аллофонов и отдельных слов в разных контекстах. Создание такой базы данных для каждого диктора является очень сложным процессом, требующим привлечения большого числа специалистов. Недавно предложен способ автоматического создания синтезатора речи по тексту, позволяющая формировать голосовую и языковую модели на основе речевых записей и их транскрипций [1]. Что касается распознавателей слитной речи, то для них при обучении системы распознавания теперь используются намного большие речевые выборки, содержащие различные голоса.

Задача изменения голоса (конверсии голоса) появилась сравнительно недавно, тем не менее развитие данного направления происходит очень активно. Целью конверсии голоса является замена личности говорящего при сохранении содержания исходного речевого сообщения. Решение данной задачи подразумевает установление соответствия между голосом исходного диктора и целевого на основе некоторого обучающего речевого материала. Большинство из всего многообразия предложенных способов решения можно разделить на несколько основных групп: использование модели Гауссовых смесей [2], масштабирование частотной шкалы [3] и использование нейронных сетей [4,5]. Последние опубликованные результаты показывают что данные подходы позволяют достигать средней разборчивости и узнаваемости целевого диктора более 75%, что, учитывая сложность задачи, является достаточно высоким показателем. Основной проблемой в конверсии голоса является несоответствие обучающих данных. Исходный и целевой дикторы даже чисто теоретически не могут говорить одинаково, вследствие этого функция конверсии страдает чрезмерным усреднением данных, что неизбежно сказывается на качестве выходного сигнала. Наиболее качественный результат конверсии достигается при использовании параллельных обучающих фраз, т.е. когда исходный и целевой дикторы произносят одни и те же обучающие фразы. Однако, наиболее интересными, исходя из практических соображений, являются решения, позволяющие выполнять обучение при по-

мощи произвольных текстовых корпусов не соответствующих друг другу [6]. Далее эти два подхода будем называть «текстозависимым» и «текстонезависимым» соответственно.

Кодирование речи является одним из самых первых направлений, сформировавшихся в цифровой обработке речевых сигналов. Наиболее эффективными здесь оказались решения, основанные на параметрическом моделировании речевого сигнала. Предложено большое количество различных моделей: линейное предсказание с различными способами возбуждения [7], кепстральный анализ [8], синусоидальное представление [9], гармоники+шум [10–12] и т.д. Современные вокодерные системы обеспечивают удовлетворительное кодирование узкополосного речевого сигнала со скоростью потока 2.4–12кбит/с и широкополосного со скоростью потока выше 8кбит/с. Не решенной в полной мере остается задача кодирования речи на сверхнизких скоростях потока менее 1.2кбит/с.

Автоматическое повышение разборчивости речи и шумоподавление представляют собой область, в которой нашли применение различные подходы: 1) универсальные методы такие как спектральное взвешивание [13–15] и обработка в подпространствах [16,17]; 2) методы основанные на особенностях слухового восприятия, такие как фильтрация в модуляционной области [18–20]; 3) методы основанные на особенностях речеобразования [21].

Все вышеперечисленные задачи относятся к одному и тому же объекту исследования (речевому сигналу), и, несмотря на все имеющиеся различия, между ними существует внутренняя взаимосвязь. Дальнейший успех в решении каждой из этих задач зависит от того, насколько удачно моделируется речь как феномен в различных его аспектах: представление процесса речеобразования, интерпретации содержания речевого сообщения (в том числе фонетического, смыслового, эмоционального) и процесса восприятия. Потому разработка способов для наиболее адекватного и универсального моделирования речевого сигнала представляется очень перспективным научным направлением.

В данной работе исследуется возможность создания параметрического описания речевого сигнала, основанного на раздельном моделировании голоса диктора и содержания речевого сообщения. Речевой сигнал представляется в виде некоторой последовательности данных, характеризующей содержание сообщения, и модели голоса диктора, при помощи которого данная последовательность преобразовывается в речь. Процесс речеобразования рассматривается как некоторая систе-

ма, позволяющая озвучивать речевые сообщения заданным голосом как показано на рисунке 1.

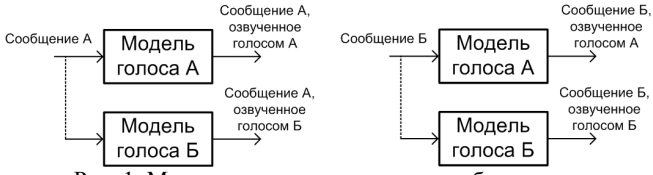

Рис. 1. Моделирование процесса речеобразования

Предполагается возможность применения модели голоса и в обратном направлении, т.е. не только для озвучивания сообщений, но и для декодирования их из речевого сигнала.

Фундаментальной проблемой является разделение параметров речевого сигнала на индивидуальные характеристики диктора и фонетическое содержимое. В чистом виде такого рода разделение естественным образом возникает в задаче синтеза речи по тексту, поскольку отдельно существует текст (содержание сообщения) и модель голоса (речевая база данных, соответствующая заданному диктору), которая используется при озвучивании этого текста. В распознавании речи решается похожая задача разделения, однако в обратном направлении: на вход поступает речевой сигнал, на выходе нужно сформировать текст сообщения.

Постановка задачи. В отличие от классических задач преобразования речь→текст и текст→речь в настоящей работе предлагается использовать несколько другую логику разделения параметров речи. Предполагается, что озвучиваемое сообщение содержит не только текстовую информацию и состоит не из последовательности фонем, а скорее из последовательности состояний речевого тракта, выполняющих универсальную фонетическую и просодическую функции для любого голоса. Таким образом, ставится задача реализации пары преобразований речь→фонемы+просодика и фонемы+просодика→речь с использованием параметрического описания голоса исходного и целевого дикторов. Постановка задачи в таком виде делает возможную область применения модели голоса очень широкой, поскольку во-первых, модель является фонетически мотивированной, однако не подразумевает явного преобразования речи в текст; во-вторых, позволяет сохранять и использовать просодику исходного речевого сообщения; в-третьих, теоретически после преобразования сохраняется возможность восстановления исходного сигнала с субъективно незначительными потерями, что невозможно достичь в результате последовательного выполнения преобразований речь→текст и текст→речь. Для

решения поставленной задачи необходимо создать параметрическую модель голоса диктора, а также соответствующие методы анализа (извлечения параметров модели из речевого сигнала) и синтеза (генерирование речевого сигнала из параметров модели). Модель голоса, предлагаемая в данной работе использует нейронную сеть, построенную по принципу автоматического кодера, реализованного в виде искусственной нейронной сети (autoencoder) [22]. В отличие от Гауссовых смесей и масштабирования частотной шкалы, нейронная сеть обладает большими возможностями, поскольку позволяет формировать более сложную функцию отображения входных и выходных характеристических векторов. Основной идеей является использование свойства кодера автоматически находить и упорядочивать похожие данные. В работе также приводятся практические результаты применения предлагаемой модели для конверсии голоса.

2. Использование модели голоса в задачах обработки речевых сигналов. Возможные области применения. В данном разделе приводятся некоторые возможные области применения параметрической модели голоса, и дается краткая характеристика вероятных преимуществ.

Конверсия голоса. При использовании параметрических моделей голоса задача конверсии голоса сводится к последовательному использованию 2-х моделей: исходного и целевого дикторов, как показано на рисунке 2. Причем, по типу обучения такая система конверсии голоса является текстонезависимой, поскольку модели голосов формируются независимо друг от друга. Такой способ обучения системы является предпочтительным, так как позволяет использовать различные речевые сообщения исходного и целевого дикторов.

Рис. 2. Использование персональной модели голоса для конверсии голоса

Кодирование речи. Разделение голоса диктора и сообщения может оказаться полезным при кодировании речи для сверхнизких скоростей передачи. Основополагающей идеей здесь служит то, что диктор не меняется (либо меняется очень редко) во время сеанса связи. Следовательно, голос диктора может быть долговременным параметром, обновляющимся значительно реже, чем параметры сообщения. При обучении системы на стороне кодера и декодера формируются универсальные базы возможных голосов. При кодировании сообщения выполняется выбор голоса из базы данных, наиболее близкого к голосу говорящего. Далее декодеру передается выделенное из речи сооб-

щение и индекс найденного голоса, как показано на рисунке 3. Учитывая, что состояния модели, используемые для описания сообщения, могут описываться набором параметров малой размерности, а так же и то, что эти параметры могут быть грубо проквантованы, данная схема кодирования может оказаться эффективной.

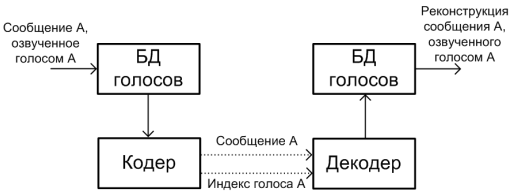

Рис. 3. Использование персональной модели голоса для кодирования речевого сигнала

Синтез речи по тексту. Использование модели голоса представляет интерес и для задачи синтеза речи по тексту. Процесс синтеза может быть организован таким образом, чтобы формирование речевого сообщения выполнялось отдельно от озвучивания его заданным голосом, как показано на рисунке 4.

Рис. 4. Использование персональной модели голоса для синтеза речи по тексту

В результате процесс добавления в базу данных синтезатора речи по тексту новых дикторов упрощается и унифицируется. При качественной автоматизации формирования модели голоса на основе речевых записей возможно создание синтезаторов речи по тексту с функцией добавления пользовательских голосов.

Шумоподавление. Модель голоса также может применяться и для очистки речи от шума – рисунок 5. Основная идея заключается в том, что речевое сообщение представляет собой последовательность состояний, которые имеют значительную продолжительность (от 10 до 50 мс), а продолжительность переходного процесса соответствующего смене состояний можно считать короткой (менее 1 мс). Таким образом, чистое сообщение должно иметь некоторую сегментную структуру. Учитывая это, процедуру фильтрации можно организовать в виде двух следующих шагов: 1) аппроксимация зашумленного сообщения наиболее вероятной последовательностью, состоящей из сегментов с ограничением минимальной длины путем решения задачи оптимизации; 2) сглаживание каждого выделенного сегмента фильтром низких частот.

Критическим фактором в данном походе является точность декодирования сообщения из входного речевого сигнала. Поскольку точность будет сильно зависеть от интенсивности и характера шума, вероятно, что данный подход может применяться только для шумов определенного типа. Интересным вопросом для исследования является также возможность подавления акустического эха и реверберации при помощи предложенной схемы.

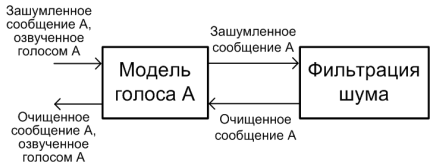

Рис. 5. Использование персональной модели голоса для шумоподавления

Повышение разборчивости. Рассмотрим возможность применения голосовой модели для повышения разборчивости речевого сигнала. Соответствующая схема обработки речи, представленная на рисунке 6, выполняет сложное преобразование просодики входящего речевого сообщения, моделирующие неторопливую и внятную речь. Используя подобную схему можно выполнять как изменение общих характеристик голоса, так и корректировку звучания отдельных фонем. Отдельной задачей является повышение разборчивости речи для людей с патологиями слуха [23]. Потеря слуха чаще всего характеризуется утратой чувствительности в высокочастотной области спектра. В результате этого в первую очередь ухудшается способность воспринимать фонемы, для которых характерны высокочастотные звуки – например [х]/[ф] и [з]/[ж]. Изменяя звучание этих отдельных фонем можно добиться повышения разборчивости. Другими возможными применениями этого подхода могут быть автоматическая коррекция акцента и искусственное расширение частотного диапазона в узкополосной телефонии.

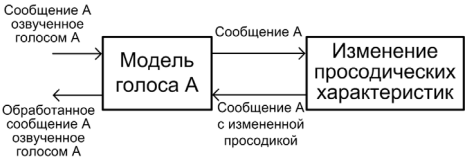

Рис. 6. Схема использования персональной модели голоса для повышения разборчивости

Во всех перечисленных выше приложениях для повышения точности декодирования сообщения дополнительно с входным рече-

вым сигналом может использоваться и его транскрипция, в случаях, когда это допустимо.

Целью данной работы является исследование возможности автоматического создания персональной модели голоса диктора по записанным образцам его голоса с транскрипцией, а также возможности практического использования такой модели для решения задачи конверсии голоса с текстонезависимым обучением.

3. Параметрическое представление речевого сигнала. Для выполнения обработки речевого сигнала необходимо выполнить его параметрическое описание, т.е. представить в виде последовательности характеристических векторов одинаковой размерности. В контексте решаемой задачи речевой сигнал удобно рассматривать как процесс имеющий квазипериодические (детерминированные) и шумовые (стохастические) составляющие [24,25]. Можно считать, что квазипериодические составляющие порождаются периодическими колебательными движениями голосовых связок и характерны для гласных (вокализованных звуков), в то время как шумовые возникают вследствие непериодических колебаний и характерны для шипящих согласных (невокализованных звуков). Моделирование процесса обеспечивается путем раздельного представления каждой из этих составляющих при помощи разных средств описания. Этот подход широко применяется в современных системах обработки речи [26,27]. В данной работе используется аналогичный способ параметрического описания, реализованный при помощи алгоритма оценки параметров периодической модели с многокомпонентным гармоническим возбуждением [28].

Речевой сигнал разбивается на перекрывающиеся фрагменты каждый из которых описывается набором параметров: спектральной огибающей, мгновенной частотой основного тона (если фрагмент вокализованный) и типом возбуждения, который может быть вокализованным, невокализованным либо смешанным.

Квазипериодическая составляющая речевого сигнала $s(n)$ представляется в виде суммы синусоид или действительной части комплексных экспонент с непрерывной амплитудой, частотой и фазой, а шумовая как случайный процесс с заданной спектральной плотностью мощности (СПМ):

$$s(n) = \sum_{p}^{P} A_p(n)\cos\varphi_k(n) + r(n) = \mathrm{Re}\left[\sum_{p}^{P} A_p(n)e^{j\varphi_p(n)}\right] + r(n),$$

где P – число синусоид (комплексных экспонент), $A_p(n)$ - мгновенная амплитуда p-ой синусоиды, $\varphi_p(n)$ - мгновенная фаза p-ой синусоиды, $r(n)$ - апериодическая составляющая. Мгновенная частота $F_p(n)$, на-

ходящаяся в интервале $[0,\pi]$ (π соответствует частоте Найквиста), является производной от мгновенной фазы. Предполагается, что амплитуда изменяется медленно, что означает ограничение частотной полосы каждой из составляющих. Используя полученные гармонические амплитуды вокализованной и СПМ невокализованной составляющих, формируется общая спектральная огибающая.

Характеристический вектор состоит из значения частоты основного тона, спектральной огибающей и признака вокализованности текущего речевого фрагмента.

Данный набор параметров, выделяется из речевого сигнала при помощи алгоритма, состоящего из следующих шагов:

– оценка мгновенной частоты основного тона при помощи устойчивого к ошибкам алгоритма слежения за мгновенной частотой основного тона IRAPT (Instantaneous Robust Algorithm for Pitch Tracking) [29];

– деформация временной оси сигнала для обеспечения стационарности частоты основного тона;

– оценка мгновенных гармонических параметров речевого сигнала с использованием ДПФ-модулированного банка фильтров – каждая гармоника основного тона вокализованной речи попадает в отдельный канал банка фильтров, где преобразуется в аналитический комплексный сигнал, из которого выделяется мгновенная амплитуда, фаза и частота;

– на основе анализа полученных значений мгновенной частоты различные области спектра классифицируются как периодические и апериодические;

– гармоники, принадлежащие периодическим областям спектра синтезируются и вычитаются из исходного сигнала;

– остаток переводится в частотную область при помощи кратковременного преобразования Фурье;

– параметры синтезированных гармоник и СПМ остатка объединяются в одну общую спектральную огибающую и переводятся в логарифмическую шкалу;

– смежные спектральные огибающие анализируются для определения способа возбуждения всего анализируемого фрагмента сигнала.

Каждая спектральная огибающая представляется в виде вектора логарифмических значений энергии равнорасположенных в шкале мелов. Для речевого сигнала с частотой дискретизации 44.1 кГц используется 100-мерный вектор. Размерность вектора определяет компромисс между качеством реконструкции сигнала и вычислительной сложностью. На основе практических экспериментов установлено, что

выбранная размерность является достаточной для реконструкции натуральной речи.

4. Модель голоса с фонетической привязкой на основе нейронной сети. Голосовая модель использует нейронную сеть, построенную по принципу автоматического кодера. Автоматический кодер представляет собой многослойную нейронную сеть, которая преобразовывает многомерные данные в коды меньшей размерности и затем восстанавливает их в первоначальном виде [22] – рисунок 7.

Понижение размерности данных широко используется в задачах классификации, связи, распознавания и др. Достаточно простым и широко применяемым способом является анализ главных компонент, выделяющий в обучающей выборке направления с максимальной дисперсией и описывающий каждый элемент выборки через координаты по этим направлениям. В работе [22] показано, что системы понижения размерности данных на основе нейронных сетей обладают гораздо более широкими возможностями, поскольку, в отличие от метода анализа главных компонент позволяют выполнять нелинейные преобразования.

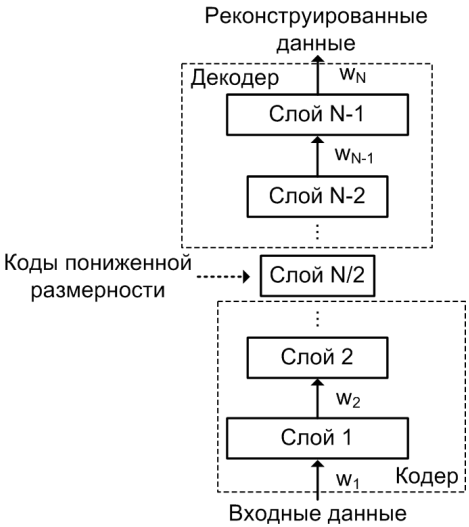

Рис. 7. Автоматический кодер на основе нейронной сети

Обе сети, кодер и декодер можно обучить вместе, уменьшая разницу между исходными данными и их реконструкцией. Частные производные всех параметров легко вычисляются, используя правило дифференцирования сложной функции, для распространения произ-

водной ошибки сперва через сеть декодирования, а затем через сеть кодирования.

В контексте конверсии голоса полезным свойством автоматического кодера является способность самостоятельно классифицировать, упорядочивать и находить похожие данные без "учителя", т.е. каких-либо заранее заданных целевых меток в обучающей выборке. Показано, что при достижении хорошей реконструкции данных, близкие коды пониженной размерности соответствуют схожим входным данным. Это может быть использовано для построения производительных ассоциативных поисковых систем [30].

В задаче конверсии голоса основной проблемой является поиск соответствия между параметрами исходного голоса и целевого. Проблема возникает из-за того, что параметры голоса выделяются на основе речевых фрагментов, соответствующим разным состояниям исходного и целевого дикторов. Каждый диктор имеет некоторое пространство возможных вариаций при произношении одних и тех же фонетических единиц, обусловленных многообразием речевых оттенков, выражением различных эмоций и интонаций. Таким образом одна и та же фонема, находясь в одном и том же фонетическом контексте может звучать по-разному. Учитывая это, очень сложно найти соответствие между состояниями разных дикторов.

Основная идея применения автоматического кодера в данной работе заключается в использовании фонетически мотивированных кодов пониженной размерности, (имеющих фонетическую интерпретацию). Предлагается организовать процесс обучения нейронной сети таким образом, чтобы коды, соответствующие одной фонеме компактно располагались в одной определенной области пространства, причем границы, примыкающих к нему областей, обеспечивали плавный переход к другим фонемам – рисунок 8.

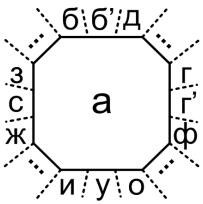

Рис. 8. Пространство кодов пониженной размерности с фонетической привязкой

Для того, чтобы обеспечить возможность непрерывного перехода из любой фонемы в любую размерность пространства равна числу используемых фонем. Расположение каждой фонемы в про-

странстве кодов фиксировано вдоль координат пространства в интервале от 0 до 1.

Использованная конфигурация искусственной нейронной сети показана на рисунке 9. Кодер выполняет функцию отображения:

$$H = (w_4 RL(w_3 RL\left(w_2 RL(w_1 X + b_1) + b_2\right) + b_3) + b_4) \otimes M$$

где X – характеристический вектор речевого сигнала, H – вектор пониженной размерности, M – вектор фонетической маски, w_{1-4} и b_{1-4} – весовые коэффициенты и смещения соответствующих сигналов сети, $\otimes$ обозначает поэлементное умножение. В сети используется кусочно-линейная функция активации $RL(x) = \max(0, x)$, поскольку показано, что она обеспечивает более эффективное внутреннее представление речевых данных по сравнению с логистической и позволяет ускорить процесс обучения [31]. На выходе кодера формируются коды пониженной размерности, на которые накладываются ограничения для того, чтобы выполнить фонетическую привязку (см. рисунок 9). Наложение ограничения выполняется путем перемножения сигнала H на фонемную маску, представляющую собой разреженную матрицу, и формируемую на основе фонетической разметки речевого корпуса.

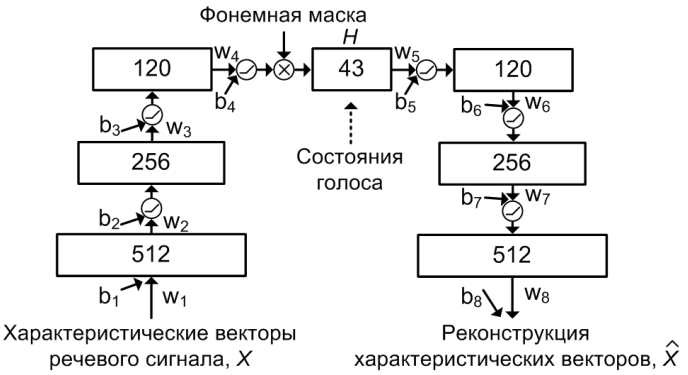

Рис. 9. Формирование состояний голоса на основе нейронной сети

Декодер выполняет реконструкцию кодов пониженной размерности в характеристические векторы $\hat{X}$. Соответствующая функция отображения имеет следующий вид:

$$\hat{X} = (w_8 RL(w_7 RL\left(w_6 RL(w_5 H + b_5) + b_6\right) + b_7) + b_8).$$

Использовалось следующее число нейронов в каждом скрытом слое нейронной сети: 512-256-120-43-120-256-512. Обучение сети включает несколько этапов, которые кратко описаны ниже.

Предварительная сегментация обучающего речевого корпуса на фонемы. Для того чтобы при обучении нейронной сети было возможно осуществить фонетическое разграничение состояний голоса, необходимо установить соответствие каждого обучающего характеристического вектора определенной фонеме. С этой целью выполняется сегментация обучающего речевого корпуса на фонемы, используя фонемную транскрипцию. Задача определения границ фонем в речевом сигнале имеет классическое решение, основанное на использовании скрытой марковской модели [32]. Для повышения точности анализа может применятся предварительная ручная (частичная либо полная) разметка.

Инициализация параметров сети и предварительное обучение. Поиск оптимальных коэффициентов многослойного кодера является сложной задачей, поскольку для того, чтобы алгоритм обратного распространения ошибки был эффективным, требуется хорошее начальное приближение. При использовании больших начальных коэффициентов процесс обучения обычно сходится к плохому локальному минимуму, использование малых коэффициентов приближает градиент в начальных слоях к нулю, что делает невозможным обучение сети с большим количеством слоев. Известен метод раздельного предварительного обучения слоев при помощи ограниченной машины Больцмана, успешно использованный в различных задачах машинного обучения [22]. В настоящей работе применяется схема предварительного обучения, основанная на частичном обнулении данных. Каждая пара матриц коэффициентов инициализируется случайными числами, и тренируется отдельно в виде нейронной сети с одним скрытым слоем. На вход сети подаются данные, часть из которых случайным образом обнуляется, на выходе сети восстанавливаются полные исходные данные. Причем обеспечивается равенство соответствующих матриц кодера и декодера $w_1 = w_8^T$, $w_2 = w_7^T$, $w_3 = w_6^T$, $w_4 = w_5^T$. Обучение выполняется при помощи алгоритма обратного распространения ошибки с накоплением градиента (momentum). При обучении матриц коэффициентов $w_4 = w_5^T$ частные производные по внутренним сигналам вычисляются с дополнитель-

ным слагаемым, обеспечивающим повышение активности в точках, обозначенных единицами в фонетической маске и понижение в точках, обозначенных нулями. Фонетическая маска формируется на основании предварительной сегментации речевого корпуса. Пространство кодов состояний голоса и маска имеют размерность 43 (равную числу фонем русского языка). Каждая фонема соответствует отдельной координате. Маска содержит единицы в координатах, соответствующих фонеме каждого вектора состояния, как показано на рисунке 10. Промежуточный переход между двумя соседними фонемами задается двумя единицами.

Рис. 10. Маска кодов пониженной размерности для осуществления фонетической привязки

Обучение системы кодера/декодера. После предварительного определения коэффициентов каждого слоя выполняется подгонка параметров всей модели при помощи модификации алгоритма обратного распространения ошибки RPROP (Resilient back PROPagation) [33]. При перемножении внутреннего сигнала на маску каждый отдельный вектор попадает либо в определенную плоскость пространства, соответствующую двум смежным фонемам, либо на координатную ось, соответствующей одной фонемы. В процессе подгонки параметров происходит уменьшение ошибки реконструкции и упорядочение характеристических векторов каждой фонемы вдоль осей пространства – рисунок 11.

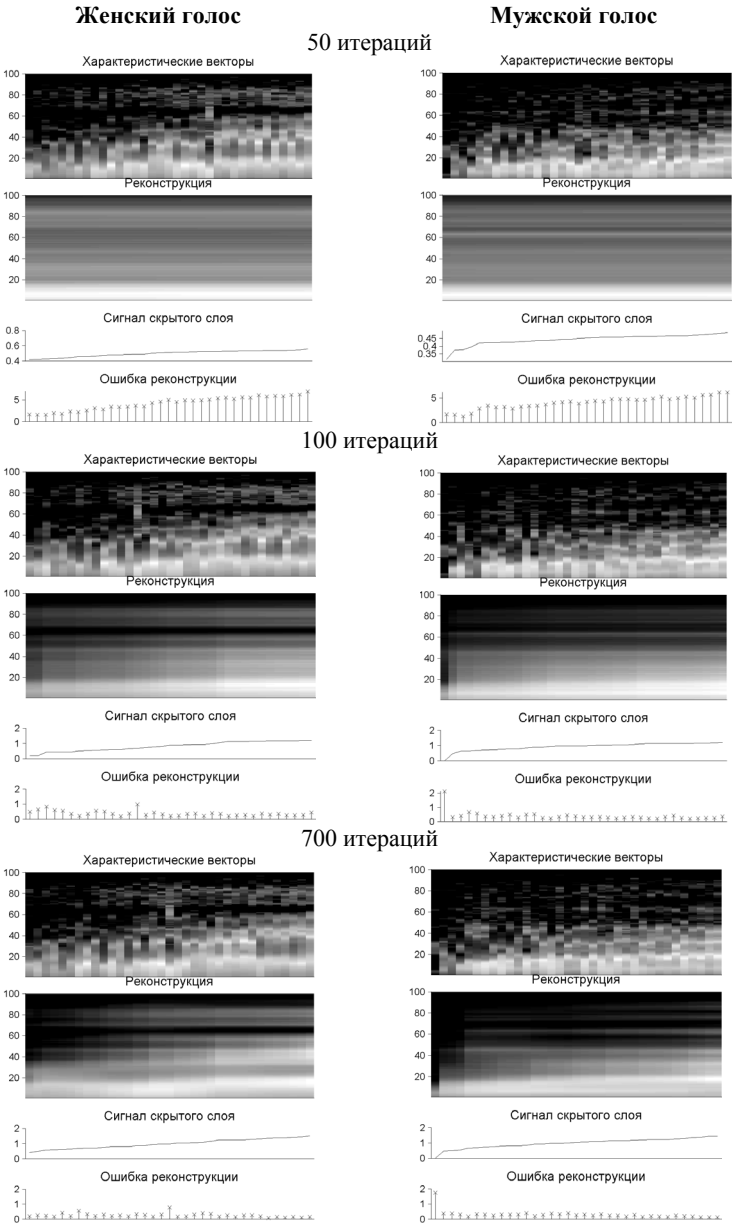

Рис. 11. Расположение характеристических векторов фонемы [а] вдоль соответствующей оси пространства кодов пониженной размерности

Процедура предварительного обучения создает достаточно хорошее начальное приближение и характеристические векторы в начале процедуры уже определенным образом упорядочены. Однако, как показано на рисунке 11, после выполнения некоторого числа итераций происходит перестановка векторов, уменьшающая ошибку реконструкции и разницу между соседними векторами. В результате обучения формируется модель голоса, которая включает модель каждой отдельной фонемы и переходов между ними, содержащихся в обучающей выборке.

5. Результаты экспериментов. Целью выполненных экспериментов является оценка практической применимости предложенной модели для решения задачи конверсии голоса с текстонезависимым обучением. Обученные модели голосов использовались для поиска соответствия между характеристическими векторами исходного и целевого дикторов. Модели обучались независимо друг от друга. Конверсия речевого сигнала выполнялась с использованием ручной фонетической разметки, в которой выделялся характерный «центральный» характеристический вектор каждой фонемы. На основании разметки автоматически определялись «переходные» характеристическим векторы, относящиеся к границам между фонемами. Соответствие между центральными и переходными характеристическими векторами исходного и целевого диктора устанавливалось следующим образом. На вход кодера исходного диктора подавался входной характеристический вектор и вычислялся вектор пониженной размерности, умноженный на фонемную маску. Из обучающей выборки целевого диктора извлекался характеристический вектор, наиболее близкий к полученному в пространстве кодов пониженной размерности. Все необходимые для синтеза характеристические векторы вычислялись путем интерполяции между конвертированными центральными и переходными векторами.

Речевая база и оценка качества конверсии. Для экспериментальной оценки качества конверсии использовалась речевая база на русском языке. База содержит широкополосную речь 6-и дикторов (3 мужчин и 3 женщины), записанную с частотой дискретизации 44,1 кГц. Для каждого из дикторов в базе содержится по 26 фраз для обучения и по 4 фразы для конверсии.

В ходе экспериментов выполнено сравнение двух методов: 1) изложенный выше метод конверсии с текстонезависимым обучением на основе автоматического кодера (далее обозначается как ‘ТН’) и метод с текстозависимым обучением на основе нейронной сети с кусочно-линейной функцией активации (далее обозначается как ‘ТЗ’) [5].

Для обоих методов использовались одинаковые алгоритмы параметризации и синтеза речевого сигнала, одинаковые обучающие

последовательности исходного и целевого дикторов, одинаковые сигналы возбуждения и одинаковые целевые контуры основного тона. Результаты конверсии оценивались субъективно в терминах узнаваемости целевого диктора и натуральности звучания реконструированной речи с использованием средних значений оценок экспертов MOS (mean opinion score). В ходе эксперимента задействовано четверо слушателей, которые ставили оценки (по шкале от 1 до 5). Усреднение оценок выполнено раздельно по четырем группам в зависимости от направления конверсии мужчина-мужчина, мужчина-женщина, женщина-мужчина, женщина-женщина (обозначенных как "мм", "мж", "жм" и "жж" соответственно) для каждого из методов конверсии. Средние оценки показаны на рисунке 12.

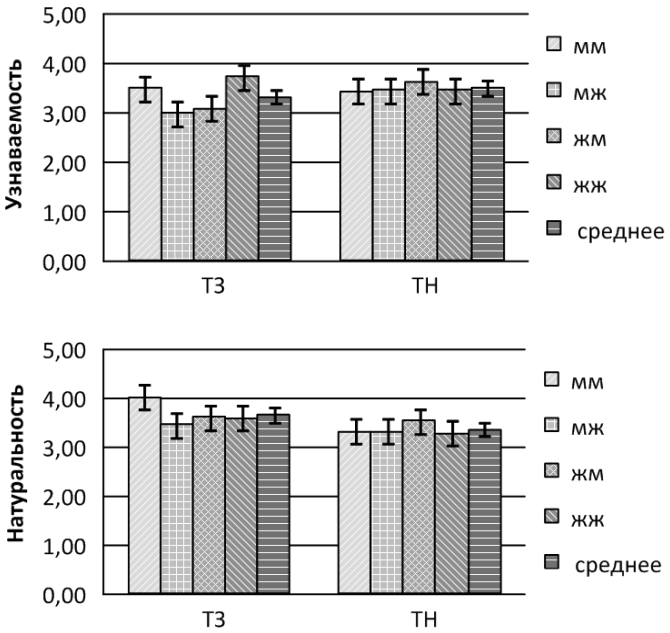

Рис. 12. Субъективная оценка узнаваемости и натуральности конвертированной речи. Средние значения оценок экспертов (доверительный интервал 95%).

На основании прослушивания и анализа оценок можно сделать вывод, что метод на основе автоматического кодера обеспечивает немного более высокую узнаваемость целевого диктора, однако несколько проигрывает по натуральности звучания. Повышение средней узнаваемости скорее всего обусловлено тем, что описанный способ позво-

ляет ослабить эффект усреднения спектральной огибающей, характерный для систем с текстозависимым обучением. Понижение натуральности обусловлено в первую очередь ошибками полуавтоматической сегментации речевого корпуса и тем, что использовалась простая модель сегментации, выделяющая только границы и центр каждой из фонем. Следует также отметить, что в методе ТЗ используется разделение параметров огибающей на высокочастотные и низкочастотные, а так же последовательность состояний диктора, генерируемых автоматически на основе текущих значений основного тона и признаков вокализованности. Таким образом, на вход системы конверсии ТЗ поступает намного больше характеристических признаков. В тоже время, необходимо учитывать, что в методе ТН использована ручная фонемная разметка, которая значительно упрощает поиск соответствия между исходными и целевыми данными.

6. Заключение. В работе исследуется возможность создания персональной модели голоса с фонетической привязкой на основе искусственной нейронной сети, построенной по принципу автоматического кодера. Приводятся результаты экспериментального применения данного подхода к решению задачи конверсии голоса с текстонезависимым обучением. Показано, что формирование унифицированных состояний в виде кодов пониженной размерности позволяет установить соответствие между различными голосами и может использоваться в системах синтеза речи по тексту и конверсии голоса. Особенностью полученной модели является относительная инвариантность к характеру произношения, что достигается за счет привязки внутренних состояний к фонетическому содержанию, что может использоваться в различных системах обработки речи, таких как системы автоматического распознавания и кодирования.

Литература

1. *Watts O., Stan A., Clark R., Mamiya Y., Giurgiu M., Yamagishi J., King S.* Unsupervised and lightly supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis // In: Proc. 8th ISCA Speech Synthesis Workshop. 2013. pp. 101–106.
2. *Toda T., Black A.W., Tokuda K.* Voice conversion based on maximum likelihood estimation of spectral parameter trajectory // IEEE Trans. Audio, Speech and Language Processing. 2007. vol. 15. no. 8. pp. 2222–2235.
3. *Godoy E., Rosec O., Chonavel T.* Spectral envelope transformation using DFW and amplitude scaling for voice conversion with parallel or nonparallel corpora // Proc. INTERSPEECH. Florence. Italy. 2011. pp. 673–676.
4. *Desai S., Black A.W., Yegnanarayana B., Prahallad B.* Spectral mapping using artificial neural networks for voice conversion // IEEE Trans. Audio, Speech and Language Processing. 2010. vol. 18. no. 5. pp. 954–964.
5. *Azarov E., Vashkevich M., Likhachov D., Petrovsky A.* Real-time Voice Conversion Using Artificial Neural Networks with Rectified Linear Units // Proc. INTERSPEECH Lyon. France. 2013. pp. 1032–1036.

6. *Erro D., Moreno A., Bonafonte A.* INCA Algorithm for Training Voice Conversion Systems From Nonparallel Corpora // IEEE Transactions on Audio, Speech, and Language Processing. 2010. vol. 18. no .5. pp. 944–953.
7. *Yeldener S., De Martin J.C., Viswanathan V.* A mixed sinusoidally excited linear prediction coder at 4 kb/s and below // Proc. ICASSP'98. 1998. vol. 2. pp. 589–592.
8. *Boucheron L.E., De Leon P.L., Sandoval S.* Low Bit-Rate Speech Coding Through Quantization of Mel-Frequency Cepstral Coefficients // IEEE Transactions on Audio, Speech, and Language Processing. 2012. vol. 20. no. 2. pp. 610–619.
9. *Etemoglu C.O., Cuperman V.* Matching pursuits sinusoidal speech coding // IEEE Transactions on Speech and Audio Processing. 2003. vol. 11, no. 5. pp. 413–424.
10. *Shlomot E., Cuperman V., Gersho A.* Hybrid coding: combined harmonic and waveform coding of speech at 4 kb/s // IEEE Transactions on Speech and Audio Processing. 2001. vol .9. no. 6. pp. 632–646.
11. *Sercov V.V., Petrovsky A.A.* An improved speech model with allowance for time-varying pitch harmonic amplitudes and frequencies in low bit-rate MBE coders // Proc. of the 6th European conference on "Speech communication and technology" (Eurospeech'99). Budapest. Hungary. 1999. pp. 1479–1482.
12. *Петровский А.А., Серков В.В.* Низкоскоростной вокодер с моделью речеобразования «гармоники+шум» // Цифровая обработка сигналов. Москва. 2002. №2. С. 61-74.
13. *Udrea R.M., Ciochina S.* Speech enhancement using spectral over-subtraction and residual noise reduction // International Symposium on Signals, Circuits and Systems. 2003. vol. 1. pp. 165–168.
14. *Петровский А.А., Борович А., Парфенюк М.* Дискретное преобразование Фурье с неравномерным частотным разрешением в перцептуальных системах редактирования шума в речи // Речевые технологии. Москва. 2008. №3. С. 16–26.
15. *Borowicz A., Parfieniuk M., Petrovsky A.A.* An application of the warped discrete Fourier transform in the perceptual speech enhancement // Speech Communication. ELSEVIER. 2006. vol. 48. pp. 1024–1036.
16. *Hansen P.S.K., Hansen P.C., Hansen S.D., Sorensen J.A.* Experimental comparison of signal subspace based noise reduction methods // IEEE International Conference on Acoustics, Speech, and Signal Processing. 1999. vol. 1, pp. 101–104.
17. *Borowich A., Petrovsky. A.* Signal subspace approach for psychoacoustically motivated speech enhancement // Speech Communication. Elsiver. 2011. vol. 53. pp. 210–219.
18. *Yu W., Brookes M.* Speech enhancement using a robust Kalman filter post-processor in the modulation domain // ICASSP–2013. 2013. pp.7457–7461.
19. *Bielawski K., Petrovsky A.A.* Speech enchancement system for hands-free telephone based on the psychoacoustically motivated filter bank with allpass frequency transformation // Proc. of the 6th European conference on "Speech communication and technology" (Eurospeech'99). Budapest. Hungary. 1999. pp.2555–2558.
20. *Петровский А.А., Башун Я.М.* Пре-процессор повышения качества зашумленной и реверберирующей речи для систем улитковой имплантации // Цифровая обработка сигналов. 2002. №2, Москва. С.48-61.
21. *Zorila T.-C., Kandida V., Stylianou Y.* Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression // In Proc. Interspeech. Portland. Oregon. 2012. pp. 635–638.
22. *Hinton G.E., Salakhutdinov R.R.* Reducing the Dimensionality of Data with Neural Networks // Science. 2006. vol. 313 no. 786. pp. 504–507.
23. *Arifianto D.* Speech intelligibility improvement of cochlear implant using release of masking // ICACSIS–2013. 2013. pp.207–211.
24. *D'Alessandro C., Yegnanarayana B., Darsinos V.* Decomposition of speech signals into deterministic and stochastic components // ICASSP-95. 1995 vol.1. pp. 760–763.

25. *Petrovsky Al., Azarov E., Petrovsky A.* Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding // Signal Processing. Fourier Related Transforms for Non-Stationary Signals. Elsiver. 2011. vol. 91. Issue 6. pp. 1489–1504.
26. *Kawaahra H., Nisimura R., Irino T., Morise M., Takahashi T., Banno B.* Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown // Proc. ICASSP. Taipei. Taiwan. 2009. pp: 3905–3908.
27. *Pantazis Y., Stylianou Y.* Improving the modeling of the noise part in the harmonic plus noise model of speech // Proc. ICASSP–2008. 2008. pp. 4609–4612.
28. *Azarov E., Vashkevich M., Petrovsky A.* Guslar: a framework for automated singing voice correction // The 39th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014). Florence. Italy. 2014. pp. 7969–7973.
29. *Azarov E., Vashkevich M., Petrovsky A.* Instantaneous pitch estimation based on RAPT framework // Proc. EUSIPCO'12. Bucharest. Romania. 2012. pp. 2787–2791.
30. *Nair V., Hinton G.E.* Rectified linear units improve restricted Boltzmann machines // Proc. ICML. Haifa. Israel. 2010.
31. *Zeiler M.D., Ranzato M., Monga R., Mao M., Yang K., Le Q.V., Nguyen P., Senior A., Vanhoucke V., Dean J., Hinton G.* On Rectified Linear Units for Speech Processing // Proc. ICASSP. Vancouver. Canada. 2013.
32. *Rabiner L.R., Juang B-H.* Fundamentals of speech recognition // Pearson Education. 1993. 507 p.
33. *Осовский С.* Нейронные сети для обработки информации // Москва: "Финансы и статистика". 2002. 344 с.

References

1. Watts O., Stan A., Clark R., Mamiya Y., Giurgiu M., Yamagishi J., King S. Unsupervised and lightly supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis. In: Proc. 8th ISCA Speech Synthesis Workshop. 2013. pp. 101–106.
2. Toda T., Black A.W., and Tokuda K. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio, Speech and Language Processing. 2007. vol. 15. no. 8. pp. 2222–2235.
3. Godoy E., Rosec O., Chonavel T. Spectral envelope transformation using DFW and amplitude scaling for voice conversion with parallel or nonparallel corpora. Proc. INTERSPEECH. Florence. Italy. 2011. pp. 673–676.
4. Desai S., Black A.W., Yegnanarayana B., Prahallad B. Spectral mapping using artificial neural networks for voice conversion. IEEE Trans. Audio, Speech and Language Processing. 2010. vol. 18. no. 5. pp. 954–964.
5. Azarov E., Vashkevich M., Likhachov D., Petrovsky A. Real-time Voice Conversion Using Artificial Neural Networks with Rectified Linear Units. Proc. INTERSPEECH Lyon. France. 2013. pp. 1032–1036.
6. Erro D., Moreno A., Bonafonte A. INCA Algorithm for Training Voice Conversion Systems From Nonparallel Corpora. IEEE Transactions on Audio, Speech, and Language Processing. 2010. vol. 18. no .5. pp. 944–953.
7. Yeldener S., De Martin J.C., Viswanathan V. A mixed sinusoidally excited linear prediction coder at 4 kb/s and below. Proc. ICASSP'98. 1998. vol. 2. pp. 589–592.
8. Boucheron L.E., De Leon P.L., Sandoval S. Low Bit-Rate Speech Coding Through Quantization of Mel-Frequency Cepstral Coefficients. IEEE Transactions on Audio, Speech, and Language Processing. 2012. vol. 20. no. 2. pp. 610–619.
9. Etemoglu C.O., Cuperman V. Matching pursuits sinusoidal speech coding. IEEE Transactions on Speech and Audio Processing. 2003. vol. 11, no. 5. pp. 413–424.

10. Shlomot E., Cuperman V., Gersho A. Hybrid coding: combined harmonic and waveform coding of speech at 4 kb/s. IEEE Transactions on Speech and Audio Processing. 2001. vol .9. no. 6. pp. 632–646.
11. Sercov V.V., Petrovsky A.A. An improved speech model with allowance for time-varying pitch harmonic amplitudes and frequencies in low bit-rate MBE coders. Proc. of the 6th European conference on “Speech communication and technology” (Eurospeech’99). Budapest. Hungary. 1999. pp. 1479–1482.
12. Petrovsky A.A, Sercov V.V. [Low – bit rate vocoder with phonation model "harmonics+noise"]. *Cifrovaya obrabotka signalov – Digital signal processing*. 2002. vol. 2. Moscow. pp. 61-74. (In Russ).
13. Udrea R.M., Ciochina S. Speech enhancement using spectral over-subtraction and residual noise reduction. International Symposium on Signals, Circuits and Systems. 2003. vol. 1. pp. 165–168.
14. Petrovsky A.A., Borowicz A., Parfieniuk M. [Discrete Fourier transform with non-uniform frequency resolution in perceptual noise reduction systems]. *Recheviye technologii – Speech technologies*. Moscow. 2008. vol. 3. pp. 16–26. (In Russ).
15. Borowicz A., Parfieniuk M., Petrovsky A.A. An application of the warped discrete Fourier transform in the perceptual speech enhancement. Speech Communication. ELSEVIER. 2006. vol. 48. pp. 1024–1036.
16. Hansen P.S.K., Hansen P.C., Hansen S.D., Sorensen J.A. Experimental comparison of signal subspace based noise reduction methods. IEEE International Conference on Acoustics, Speech, and Signal Processing. 1999. vol. 1, pp. 101–104.
17. Borowich A., Petrovsky. A. Signal subspace approach for psychoacoustically motivated speech enhancement. Speech Communication. Elsiver. 2011. vol. 53. pp. 210–219.
18. Yu W., Brookes M. Speech enhancement using a robust Kalman filter post-processor in the modulation domain. ICASSP–2013. 2013. pp.7457–7461.
19. Bielawski K., Petrovsky A.A. Speech enchancement system for hands-free telephone based on the psychoacoustically motivated filter bank with allpass frequency transformation. Proc. of the 6th European conference on “Speech communication and technology” (Eurospeech’99). Budapest. Hungary. 1999. pp.2555–2558.
20. Petrovsky A.A., Bashun Y.M. [Pre-processor for quality improvement of noisy and reverberant speech for hearing implants]. *Cifrovaya obrabotka signalov – Digital signal processing*. Moscow. 2002. vol. 2. pp. 48-61. (In Russ).
21. Zorila T.-C., Kandida V., Stylianou Y. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In Proc. Interspeech. Portland. Oregon. 2012. pp. 635–638.
22. Hinton G.E., Salakhutdinov R.R. Reducing the Dimensionality of Data with Neural Networks. Science. 2006. vol. 313 no. 786. pp. 504–507.
23. Arifianto D. Speech intelligibility improvement of cochlear implant using release of masking. ICACSIS–2013. 2013. pp.207–211.
24. D'Alessandro C., Yegnanarayana B., Darsinos V. Decomposition of speech signals into deterministic and stochastic components. ICASSP-95. 1995 vol.1. pp. 760–763.
25. Petrovsky Al., Azarov E., Petrovsky A. Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding. Signal Processing. Fourier Related Transforms for Non-Stationary Signals. Elsiver. 2011. vol. 91. Issue 6. pp. 1489–1504.
26. Kawaahra H., Nisimura R., Irino T., Morise M., Takahashi T., Banno B. Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown. Proc. ICASSP. Taipei. Taiwan. 2009. pp: 3905–3908.
27. Pantazis Y., Stylianou Y. Improving the modeling of the noise part in the harmonic plus noise model of speech. Proc. ICASSP–2008. 2008. pp. 4609–4612.

28. Azarov E., Vashkevich M., Petrovsky A. Guslar: a framework for automated singing voice correction. The 39th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014). Florence. Italy. 2014. pp. 7969–7973.
29. Azarov E., Vashkevich M., Petrovsky A. Instantaneous pitch estimation based on RAPT framework. Proc. EUSIPCO'12. Bucharest. Romania. 2012. pp. 2787–2791.
30. Nair V., Hinton G.E. Rectified linear units improve restricted Boltzmann machines. Proc. ICML. Haifa. Israel. 2010.
31. Zeiler M.D., Ranzato M., Monga R., Mao M., Yang K., Le Q.V., Nguyen P., Senior A., Vanhoucke V., Dean J., Hinton G. On Rectified Linear Units for Speech Processing. Proc. ICASSP. Vancouver. Canada. 2013.
32. Rabiner L.R., Juang B-H. Fundamentals of speech recognition. Pearson Education. 1993. 507 p.
33. Osovsky S. Nejronnye seti dlja obrabotki informacii [Neural networks for information processing]. Moskva: "Finansyi i statistika". 2002. 344 p. (In Russ).

Азаров Илья Сергеевич — к-т техн. наук. доцент кафедры электронных вычислительных средств БГУИР. Область научных интересов: цифровая обработка речевых сигналов. Число научных публикаций — 42. azarov@bsuir.by, www.bsuir.by; БГУИР, ул. П. Бровки 6, г. Минск, 220013, РБ; р.т. +375 (17) 293-8805.

Azarov Elias — Ph.D., associate professor of computer engineering department, BSUIR. Research interests: digital speech processing. The number of publications — 40. azarov@bsuir.by, www.bsuir.by; BSUIR, 6, P.Brovky str., 220013, Minsk, RB; office phone +375 (17) 293-8805.

Петровский Александр Александрович — д-р техн. наук, профессор, заведующий кафедрой электронных вычислительных средств БГУИР. Область научных интересов: цифровая обработка сигналов. Число научных публикаций — более 600. palex@bsuir.by, www.bsuir.by; БГУИР, ул. П. Бровки 6, г. Минск, 220013, РБ; р.т. +375 (17) 293-2340.

Petrovsky Alexander — Ph.D., Dr. Sci., professor, head of computer engineering department, BSUIR. Research interests: digital speech processing. The number of publications — more than 600. palex@bsuir.by, www.bsuir.by; BSUIR, 6, P.Brovky str., 220013, Minsk, RB; office phone +375 (17) 293-2340.

РЕФЕРАТ

Азаров И.С., Петровский А.А. **Формирование персональной модели голоса диктора с универсальным фонетическим пространством признаков на основе искусственной нейронной сети.**

В работе исследуется возможность формирования модели голоса заданного диктора на основе записей образцов его голоса с транскрипцией. Необходимость решения данной задачи возникает во многих речевых приложениях таких как конверсия голоса, коррекция акцента, синтез речи по тексту, кодирование и др.

Разработанная схема моделирования голоса основывается на нейронной сети с кусочно-линейной функцией активации, построенной по принципу автоматического кодера. В результате преобразования характеристических векторов речи сетью, они представляются в виде кодов пониженной размерности с фонетической привязкой.

В практической части работы предлагаются результаты экспериментов применения голосовой модели к задаче конверсии голоса с текстонезависимым обучением.

SUMMARY

Azarov E., Petrovsky A. **Training personal voice model of a speaker with unified phonetic space of features using artificial neural network.**

The paper investigates possibility of training a personal voice model for given speaker using transcribed speech samples. Solution to the problem is required in many speech processing applications such as voice conversion, accent correction, text-to-speech synthesis, coding and other.

The proposed voice modeling scheme is based on neural network with rectified-linear units designed as deep autoencoder. Characteristic speech vectors are transformed by the network into low-dimensional codes with phonetic alignment.

In the practical part of the work some experimental results are given, that show applicability of the model to voice conversion using non-parallel training sets.