OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

РАЗДЕЛЬНОЕ МОДЕЛИРОВАНИЕ РЕЧЕВОГО СООБЩЕНИЯ В ВИДЕ ГОЛОСОВЫХ, ФОНЕТИЧЕСКИХ И ПРОСОДИЧЕСКИХ ПАРАМЕТРОВ

Азаров И.С., Петровский А.А.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь

azarov@bsuir.by

palex@bsuir.by

В работе исследуется возможность раздельного моделирования речевого сообщения (речевого сигнала) путем разделения его на голосовую, фонетическую и просодическую составляющие. В работе предлагается практический способ построения модели сообщения и приводятся результаты ее экспериментального 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: модель речевого сообщения, фонетика, просодика, искусственные нейронные сети

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. Для качественного синтеза речи по тексту сегодня вместо аллофонного синтеза применяется корпусный, использующий десятки часов речевого материала. Создание такой базы данных для каждого диктора является очень сложным процессом, требующим привлечения большого числа специалистов. 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. Основной проблемой в конверсии голоса является несоответствие обучающих данных. Самый качественный результат конверсии достигается при использовании параллельных обучающих фраз, т.е. когда исходный и целевой дикторы произносят одни и те же обучающие фразы. Однако, более практичными, являются решения, позволяющие выполнять обучение при помощи произвольных текстовых корпусов не соответствующих друг другу. Далее эти два подхода будем называть «текстозависимым» и «текстонезависимым» соответственно. Последние опубликованные результаты показывают что возможно достигать средней разборчивости и узнаваемости целевого диктора более 75%, что, учитывая сложность задачи, является достаточно высоким показателем.

Кодирование речи является одним из самых первых направлений в цифровой обработке речевых сигналов. Наиболее эффективными здесь являются решения, основанные на параметрическом моделировании речевого сигнала. 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.

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 перспективным научным направлением.

В данной работе исследуется возможность создания параметрического описания речевого сигнала, основанного на раздельном моделировании голоса диктора и содержания речевого сообщения. Фундаментальной проблемой является разделение параметров речевого сигнала на индивидуальные характеристики диктора, а также фонетическое и просодическое содержимое. Похожее 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. 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. Модель речевого сообщения, предлагаемая в данной работе использует нейронную сеть, построенную по принципу автоматического кодера, реализованного в виде искусственной нейронной сети (autoencoder) [Hinton G.E., 2006]. В работе также приводятся практические результаты применения предлагаемой модели для конверсии голоса.

1. 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 [D'Alessandro C., 1995]. 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 [Kawaahra H., 2009]. В данной работе используется аналогичный способ параметрического описания, реализованный при помощи алгоритма оценки параметров периодической модели с многокомпонентным гармоническим возбуждением [Azarov E., 2014].

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):

$$s(n) = \sum_{p}^{P} A_p(n)\cos\varphi_k(n) + r(n)$$

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 $F_p(n)$, находящаяся 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.

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) [Azarov E., 2012];
- 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.

2. Модель речевого сообщения с фонетической привязкой на основе нейронной сети

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 [Hinton G.E., 2006].

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. 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. 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 производной 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 [Nair V., 2010].

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 переход к другим фонемам. Для того, чтобы обеспечить возможность непрерывного перехода из любой фонемы в любую размерность пространства равна числу используемых фонем. Расположение каждой фонемы в пространстве кодов фиксировано вдоль координат пространства в интервале от 0 до 1.

Использованная конфигурация искусственной нейронной сети показана на рисунке 1. Кодер выполняет функцию отображения

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

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 [Zeiler M.D., 2013]. На выходе кодера формируются коды пониженной размерности, на которые накладываются ограничения для того, чтобы выполнить фонетическую привязку. Наложение ограничения выполняется путем перемножения сигнала H на фонемную маску, представляющую собой разреженную матрицу, и формируемую на основе фонетической разметки речевого корпуса.

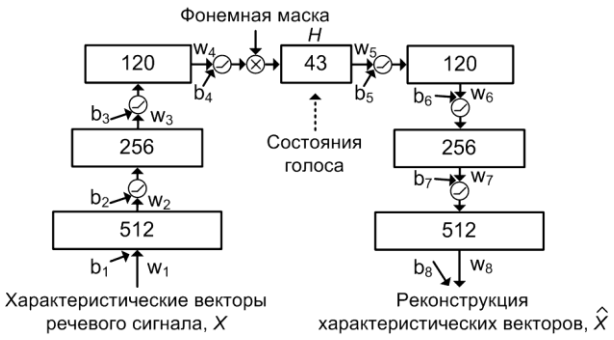

Рисунок 1 - Формирование состояний голоса на основе нейронной сети

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

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

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.

2.1. 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 [Rabiner L.R., 1993]. Для повышения точности анализа может применятся предварительная ручная (частичная либо полная) разметка.

2.2. 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 [Hinton G.E., 2006]. 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 дополнительным 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 2. Промежуточный переход между двумя соседними фонемами задается двумя единицами.

2.3. 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) [Осовский С.]. 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 2.

Рисунок 2 - Маска кодов пониженной размерности для осуществления фонетической привязки

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 сообщения, которая включает модель голоса диктора и модель каждой отдельной фонемы и возможных переходов между ними, содержащихся в обучающей выборке.

3. 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.

3.1. 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’) [Azarov E., 2013].

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 3.

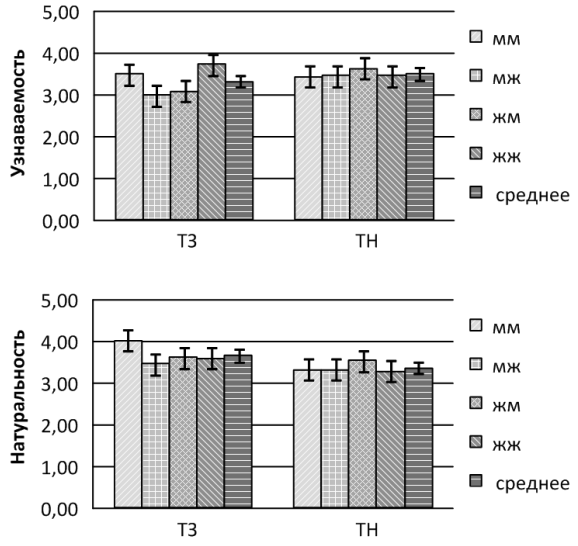

Рисунок 3 - Субъективная оценка узнаваемости и натуральности конвертированной речи. Средние значения оценок экспертов (доверительный интервал 95%).

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, позволяет 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.

Заключение

В работе исследуется возможность создания модели речевого сообщения 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.

Библиографический список

[Hinton G.E., 2006] Hinton G.E., Salakhutdinov R.R. Reducing the Dimensionality of Data with Neural Networks, Science // Vol. 313 no. 5786 pp. 504-507, 28 July, 2006.

[D'Alessandro C., 1995] D'Alessandro C., Yegnanarayana B., Darsinos V. Decomposition of speech signals into deterministic and stochastic components // ICASSP-95., vol.1, pp.760–763, 9–12 May 1995.

[Kawaahra H., 2009] 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, April 2009.

[Azarov E., 2014] 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, May 2014. – P. 7969–7973.

[Azarov E., 2012] Azarov E., Vashkevich M., Petrovsky A. Instantaneous pitch estimation based on RAPT framework // Proc. EUSIPCO'12, Bucharest, Romania, Aug. 2012, pp. 2787–2791.

[Nair V., 2010] Nair V., Hinton G.E. Rectified linear units improve restricted Boltzmann machines // Proc. ICML, Haifa, Israel, June 2010.

[Zeiler M.D., 2013] 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, May 2013.

[Rabiner L.R., 1993] Rabiner L.R., Juang B-H. Fundamentals of speech recognition // Pearson Education, 1993, P. 507.

[Осовский С.] Осовский С. Нейронные сети для обработки информации // Москва: "Финансы и статистика", 2002, 344 с.

[Azarov E., 2013] Azarov E., Vashkevich M., Likhachov D., Petrovsky A. Real-time Voice Conversion Using Artificial Neural Networks with Rectified Linear Units // Proc. INTERSPEECH Lyon, France, Aug. 2013, pp. 1032–1036.

SEPARATE MODELING OF SPEECH USING VOICE, PHONETICAL AND PROSODIC CHARACTERISTICS

Azarov E., Petrovsky A.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

azarov@bsuir.by

palex@bsuir.by

The paper investigates possibility of speech modeling using separate voice, phonetical and prosodic components. The paper presents a practical way of building such 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.