Конверсия голоса на основе множественной регрессионной функции отображения и метода спектрального взвешивания

Петровский А.А., *доктор технических наук*

Захарьев В.А., *аспирант*

В статье рассматриваются вопросы развития методов и моделей конверсии голоса. Приводится обзор и анализ наиболее часто используемых подходов на основе статистических методов. На их базе в работе предлагаются расширенные модели с большим количеством степеней свободы и факторов, которые учитываются при построении функции конверсии, а также адекватностью модели по отношению к характеристикам речевого сигнала. В статье предлагается усовершенствованный статистический метод на основе множественной регрессии, а также метод на базе спектрального взвешивания. Эффективность работы предложенных методов подтверждается объективными тестами, результаты которых приведены в экспериментальной части статьи.

• конверсия голоса • гауссовы смеси • множественная регрессионная модель • спектральное взвешивание

The paper deals with the development of methods and models for voice conversion. Provides an overview and analysis of the most commonly used approaches based on statistical methods. On their basis in the article offers enhanced model with a large number of degrees of freedom and the factors that are taken into account when constructing the function of conversion, as well as the adequacy of the model in relation to the characteristics of the speech signal. The paper proposes an improved method based on a statistical multiple regression, and the method based on the spectral weighting. The effectiveness of the proposed methods is confirmed by objective tests , the results of which are given in the experimental part of the article.

• voice conversion • Gaussian mixture model • Multivariate regression model • spectral weighting

Введение

С каждым годом к системам человеко-машинного взаимодействия предъявляются всё более жёсткие требования к их качественным характеристикам: натуральности, воспроизведению персонализированных свойств голоса, а также мультимодальности речевых интерфейсов и систем мультимедиа. Данный факт обусловил формирование и развитие одного из наиболее молодых и бурно развивающихся в настоящее время направлений речевых исследований – конверсии голоса. Конверсия голоса является технологией обработки речевого сигнала, позволяющей реализовать процесс трансформации параметров голоса, характеризующих речь исходного диктора, в параметры целевого [1, 2]. Данная технология находит своё широкое применение в области создания многодикторных систем синтеза речи по тексту, биометрических систем фоноскопической экспертизы в криминалистике, индустрии развлечений, систем восстановления голоса [3].

Объектами конверсии голоса, как технологии обработки сигналов, являются информационные характеристики диктора, проявляющиеся в речевом сигнале через изменение его акустических параметров. Если постулироваться акустической моделью речеобразования «источник – фильтр», то можно выделить два вида таких параметров: артикуляторные и просодические характеристики. К артикуляторным можно отнести тембральные свойства голоса, задающие спектральную окраску фонем, проявляющиеся в виде изменения спектральной огибающей сигнала. Просодические характеристики – это совокупность физических параметров речевого сигнала, посредством которых реализуются интонация и ударения в речи. К ним относятся: мелодика – движение частоты основного тона, ритмика – текущее изменение длительности звуков и пауз, энергетика – текущее изменение силы звука [4].

Данная статья посвящена рассмотрению вопросов анализа и улучшения существующих моделей конверсии тембральных характеристик диктора как наиболее существенной составляющей данной технологии. Поскольку именно адекватностью и сложностью данной модели конверсии в большей степени определяется точность, с которой могут быть аппроксимированы параметры исходного диктора параметрами целевого, а следовательно, и перцептивное качество сконвертированной речи.

1. Конверсия голоса на основе статистических методов

Центральной задачей конверсии голоса является поиск функции конверсии голоса, позволяющей выполнить оптимальное отображение вектора параметров исходного диктора на каждом фрейме анализа в параметры целевого диктора [5]. В качестве такого критерия оптимальности, как правило, выступает минимум расстояния между векторами в пространстве акустических параметров пары дикторов. Под пространством акустических параметров диктора будем понимать всю совокупность векторов параметров, характеризующих спектральную огибающую, полученную в результате анализа и параметризации речевого сигнала на основе выбранного метода его представления. Таким образом, мы *постулируемся тем*, что на вход системы поступает последовательность характеристических векторов, с помощью которых закодирована спектральная огибающая диктора, одновременно абстрагировавшись от конкретной модели сигнала или метода его представления, которая в данном случае может быть произвольной.

Функция конверсии строится на основе представлений о структуре таких акустических пространств и их взаимосвязях друг с другом, опирается на различные модели их мате-

матического описания [6–8], позволяющие выразить такие представления, а также решить в процессе поиска функции конверсии две основные задачи. Во-первых, осуществить разделение пространства на характерные участки (кластеры), соответствующие определённым акустическим событиям – характерному изменению состояния параметров сигнала – как правило, фонемам звуков речи, объединённых в кластеры по месту или способу образования. Во-вторых, задавшись возможным характером взаимосвязи между пространствами дикторов, определить параметры функции конверсии на основе выбранного метода кластеризации и характеристик самих кластеров, полученных по результатам первого этапа.

В системах конверсии голоса для достижения поставленного результата используются различные методы кластеризации и функции конверсии на базе статистических методов, линейной алгебры, эвристических и методов обработки сигналов. Результаты применения и примеры реализации данных методов можно также найти как в зарубежной [9], так и отечественной литературе [10–12]. В настоящий момент времени одним из самых распространённых и широко используемых методов, доказавших свою эффективность применения, является модель на основе множественных гауссовых смесей [13].

1.1. Модель множественных гауссовых смесей

Множественные гауссовы смеси — это вероятностная модель, которая позволяет представить акустическое пространство одного или более дикторов набором перекрывающихся классов, с возможностями определения характеристик модели по методу «обучение без учителя» [14]. Класс или компонента смеси отражают некоторые особенности речи говорящего диктора, непосредственно связанные с фонетическими событиями. Выбор количества классов представляет собой классическую задачу на разрешение технического противоречия между точностью представления пространства и сложностью модели. Во многих ситуациях количество выбирается эмпирическим путём, в зависимости от требуемой точности и желаемой детализации фонетических событий. Однако в общем случае является отдельной самостоятельной задачей, требующей детального рассмотрения. Описание каждого класса выполняется с использованием параметров нормального распределения: вектором средних значений класса и ковариационной матрицей, определяющей форму дисперсионного рассеивания векторов параметров вокруг среднего вектора в пределах смеси. Размерность среднего вектора и ковариационной квадратной матрицы смеси соответствует размерности входного вектора параметров сигнала. Поэтому при выборе модели представления сигнала необходимо стараться выбирать такие из них, которые либо обладают невысокой размерностью выходного вектора параметров, либо позволяют применить методы снижения размерности. Целиком акустическое пространство параметров диктора описывается набором классов или смесью из гауссовых компонент, каждая из которых имеет свой весовой коэффициент, средний вектор и ковариационную матрицу.

После нахождения показателей модели она сразу же может использоваться для классификации входных векторов акустических параметров с использованием правила Байеса. Классификация носит вероятностный и непрерывный характер.

Функция плотности вероятности модели на основе множественных гауссовых смесей представляет собой взвешенную сумму Q гауссовых компонент и определяется следующим выражением:

$$p(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{q=1}^{Q} \alpha_q N(\boldsymbol{x}|\Theta_q),$$

где $x = [x_o, x_o,...x_{p-1}]^T$ – случайный вектор размерности p, $N_q(x|\Theta_q)$ – плотность вероятности компоненты смеси, а α_q – её весовой коэффициент. Каждая из компонент представляет собой функцию плотности вероятности размерности *p*:

$$N_q(\boldsymbol{x}|\Theta_q) = \frac{1}{(2\pi)^{p/2}} |\boldsymbol{\Sigma}_q|^{-1/2} \mathbf{e}^{\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q (\boldsymbol{x}-\boldsymbol{\mu}_q)\right]},$$

где μ_q – вектор математических ожиданий класса размерностью p и $\sum_q$– ковариационная матрица многомерного распределения Гаусса размерностью pxp. Скалярные веса смесей α_q принимают значения больше нуля $\alpha_q \geq 0$, $\forall q = 1,...,Q$, а их сумма равна единице $\sum_{q=1}^{0} \alpha_q = 1$. Таким образом, полное параметрическое представление модели множественных гауссовых смесей, описывающей акустическое пространство диктора, включает в себя $\Theta = \{\alpha_q, \mu_q, \sum_q\}$ характеристики для q =1,...,Q компонент.

1.2. Функция конверсии на основе регрессии первого порядка

Для описания акустического пространства диктора на основе модели множественных гауссовых смесей в работе [13] были предложены функции конверсии, основанные на мягкой классификации. Результаты конверсии на основе данных функций выгодно отличались от конверсии на основе подходов с жёсткой кластеризацией пространства параметров, например, векторного квантования, поскольку позволяли избежать возникновения артефактов в выходном речевом сигнале. Первоначально формула функции конверсии для множественной смеси была получена на основе использования регрессионной модели первого порядка для однокомпонентного случая. Если постулироваться тем, что характеристики векторов ИД и ЦД имеют нормальное распределение, минимум среднеквадратичной ошибки для преобразованного вектора определяется регрессионным уравнением первого порядка вида [15]:

$$\mathbb{E}[\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{x}_n] = \mathbf{v} + \boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \quad (1)$$

где Γ– кроссковариационная матрица для векторов ИД и ЦД, а **v** – вектор средних значений для ЦД:

$$\mathbf{v} = E[y],$$

$$\Gamma = E[(y - \mathbf{v})(x - \mu)^T].$$

Полученный результат был расширен для случая множественных смесей путём умножения каждой компоненты, определяемой выражением (1), на весовой множитель соответствующей смеси, который определяется условной вероятностью принадлежности поступающего на вход функции вектора x_n к классу w_q, описываемому данной смесью. Тогда общая форма функции конверсии может быть представлена как:

$$F(\boldsymbol{x}) = \sum_{q=1}^{Q} p_q(\boldsymbol{x})[\mathbf{v}_q + \boldsymbol{\Gamma}_q \boldsymbol{\Sigma}_q^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_q)], \quad (2)$$

где $p_q(x)$– апостериорная вероятность того, что вектор x принадлежит q-й гауссовой компоненте.

Параметры $[v_q, \Gamma_q]$ вычисляются с применением методов среднеквадратической оптимизации с целью минимизации ошибки преобразования между сконвертированными и целевыми данными на тренировочной выборке:

$$\varepsilon_{mse} = \mathbb{E}[\|\boldsymbol{y} - F(\boldsymbol{x})\|^2].$$

В работе [16] данная стратегия конверсии была применена к параметризации на основе кепстральных коэффициентов. Сравнивались два подхода к построению данной функции в зависимости от типа кроссковариационной матрицы Γ_q: для случая полной ковариационной матрицы Γ и диагональной ковариационной матрицы. Было показано, что в результате больших временных и ресурсных затрат, возникающих при использовании не диагональных матриц при их инверсии, более предпочтительными являются диагональные кроссковариационные матрицы.

В работе [17] метод поиска параметров функции конверсии (2) на базе наименьших квадратов был расширен предположением о том, что возможно построение модели множественных гауссовых смесей для совместного пространства векторов исходного и целевого дикторов $z = [x^T, y^T]^T$ с целью возможностей описания совместной плотности вероятности $p(x,y)$. В данном случае функция конверсии, которая минимизирует среднеквадратическую ошибку между сконвертированным и целевым векторами, является регрессионной функцией y от x:

$$F(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = \int \boldsymbol{y} p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y} = \sum_{q=1}^{Q} p_q(\boldsymbol{x})\, \hat{y}_q, \tag{3}$$

$$p_q(\boldsymbol{x}) = \frac{\alpha_q N(\boldsymbol{x}|\boldsymbol{\mu}_q^{xx}, \boldsymbol{\Sigma}_q^{xx})}{\sum_{j=1}^{Q} \alpha_j N(\boldsymbol{x}|\boldsymbol{\mu}_q^{xx}, \boldsymbol{\Sigma}_q^{xx})}, \tag{3a}$$

$$\hat{\boldsymbol{y}}_q = \boldsymbol{\mu}_q^y + \boldsymbol{\Sigma}_q^{yx} \boldsymbol{\Sigma}_q^{xx^{-1}} (\boldsymbol{x} - \boldsymbol{\mu}_q^x)$$

$$\boldsymbol{\Sigma}_q = \begin{bmatrix} \boldsymbol{\Sigma}_q^{xx} & \boldsymbol{\Sigma}_q^{xy} \\ \boldsymbol{\Sigma}_q^{yx} & \boldsymbol{\Sigma}_q^{yy} \end{bmatrix} \boldsymbol{\mu}_q = \begin{bmatrix} \boldsymbol{\mu}_q^x \\ \boldsymbol{\mu}_q^y \end{bmatrix}.$$

При таком подходе к описанию совместного акустического пространства признаков отсутствует необходимость инверсии больших часто плохо определённых матриц, поскольку все параметры регрессионной функции отображения $\{\mu_q, \Sigma_q\}$, необходимой для реализации конверсии, рассчитываются уже во время обучения модели гауссовых смесей. Метод на основе совместной плотности вероятности не делает никаких предположений о характере распределения значений векторов исходного и целевого дикторов. В теории, моделирование совместной плотности должно привести к более справедливому распределению компонентов смеси для регрессионной модели. В работах показано [18], что модели на базе метода наименьших квадратов и совместной плотности распределения приводят к одинаковым результатам. Это означает, что параметры целевого диктора имеют сходное распределение с целевыми по отношению к дисперсии. Однако в зависимо-

сти от количества обучающих данных регрессионная функция на основе модели совестных множественных гауссовых смесей оказывается лучше благодаря меньшему значению вычислительной ошибки. Из чего был сделан вывод, что данная модель имеет запас устойчивости при меньшем количестве обучающих данных.

Общей проблемой обоих подходов на основе модели множественных гауссовых смесей является расширение ширины полосы пропускания формант в результате локального усреднения параметров спектральной огибающей на выходе функции конверсии. Перцептуально это воспринимается как эффект размытия в сконвертированной речи. Данный эффект хорошо виден по отношению к четвёртой форманте сконвертированной огибающей (*рис. 1*).

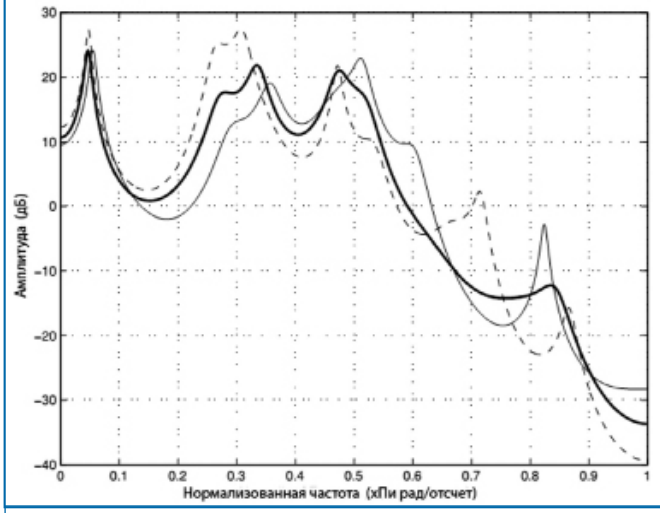

Рис. 1. Спектральные огибающие на фрейме сигнала: пунктирная линия – исходного диктора; тонкая линия – целевого диктора; толстая линия – результат конверсии

Также необходимо отметить относительно невысокую гибкость модели, поскольку функция отображения имеет всего одну степень свободы. Из анализа вида функции отображения видно, что данная функция является простейшей регрессионной функцией первого порядка вида $y = \beta_0 - \beta_1 x$, устанавливающей зависимость между одним предиктором и одной критериальной (зависимой) переменной, вследствие этого имеет ограниченные возможности предсказания. Поэтому в следующем разделе предлагается рассмотреть имеющиеся возможности по улучшению модели за счёт её усложнения и введения в неё новых факторов.

1.3. Функция конверсии на основе множественной регрессии

Эксперименты над системой, построенной на базе представленной выше функции, показали хорошие результаты. Однако было установлено, что при проведении тестирования по «слепым методикам» эксперты лишь в половине случаев принимали решение, что результирующий голос похож на голос целевого диктора, а в половине – на голос исходного. Как уже было сказано выше, данная модель является весьма ограниченной, поскольку рассматривает последовательность векторов обучения как простой набор элементов, для которых статистические связи присутствуют лишь для одной пары в каждый i-*й* момент времени (*рис. 2а*).

Выражение (3) можно переписать в упрощенном виде:

$$\boldsymbol{y}_i = \sum_{q=1}^{Q} p_q(\boldsymbol{x}_i)\left[\boldsymbol{\nu}_q + \boldsymbol{\Phi}_q \overline{\boldsymbol{x}}_i^q\right] + \boldsymbol{\varepsilon}_i, \qquad (4)$$

$$\boldsymbol{\nu}_q = \boldsymbol{\mu}_q^y, \quad \boldsymbol{\Phi}_q = \boldsymbol{\Sigma}_q^{yx}, \quad \overline{\boldsymbol{x}}_i^q = \boldsymbol{\Sigma}_q^{xx\,-1}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_q^x\right),$$

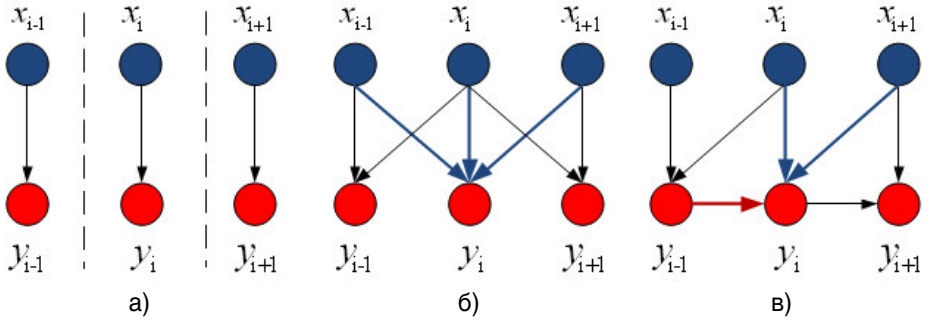

Рис. 2. Виды зависимостей между парами векторов обучающей последовательности: а) независимая модель; б) эргодическая модель; в) Марковский процесс

где ε_i – величина невязки между результатом конверсии и целевым вектором параметров, i = 1 2,...,T – момент времени, который соответствует номеру вектора в последовательности. Одним из возможных способов расширения функции конверсии является введение в модель дополнительных переменных, использующих контекстную информацию из соседних с i-м вектором элементов. Поскольку обучающая последовательность векторов параметров речевого сигнала обладает некоторой эргодичностью, была выдвинута гипотеза о том, что параметры контекстных векторов также могут коррелировать с i-м вектором целевого диктора. В зависимости от типа такой связи было предложено два подхода к расширению функции модели. Первая условно была названа эргодической моделью, она представлена на *рис. 2б* и имеет вид:

$$\boldsymbol{y}_i = \sum_{q=1}^{Q} p_q(\boldsymbol{x}_i, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}) \left[\boldsymbol{\nu}_q + \boldsymbol{\Phi}_q \bar{\boldsymbol{x}}_i^q + \boldsymbol{\Psi}_q \bar{\boldsymbol{x}}_{i-1}^q + \boldsymbol{\Omega}_q \bar{\boldsymbol{x}}_{i+1}^q \right],$$

Данная функция отображения показала лучшие результаты, чем функция (4). Далее это будет видно из экспериментов. Второй подход предлагает учитывать зависимость в последовательности векторов не только для исходного, но и для целевого диктора, придавая тем самым последовательности свойства Марковского процесса. Предлагаемая схема приведена на *рис. 2в*. Таким образом, если элементы обучающей выборки целевого диктора условно считать состояниями модели, то регрессия учитывает следующие состояния:

$$\boldsymbol{y}_i = \sum_{q=1}^{Q} p_q(\boldsymbol{x}_i, \boldsymbol{y}_{i-1}, \boldsymbol{x}_{i+1}) \left[\boldsymbol{\nu}_q + \boldsymbol{\Phi}_q \bar{\boldsymbol{x}}_i^q + \boldsymbol{\Psi}_q \bar{\boldsymbol{y}}_{i-1}^q + \boldsymbol{\Omega}_q \bar{\boldsymbol{x}}_{i+1}^q \right], \quad (5)$$

Для определения параметров множественной регрессионной функции метод на основе совместной плотности вероятности применён быть не может. Покажем, как осуществляется поиск коэффициентов данной модели на основе общего метода на базе наименьших квадратов. Представим выражение (5) в матричном виде:

$$[\mathbf{P} \vdots \mathbf{B} \vdots \mathbf{C} \vdots \mathbf{D}] \cdot \begin{bmatrix} \boldsymbol{\nu} \\ \cdots \\ \boldsymbol{\Phi} \\ \cdots \\ \boldsymbol{\Psi} \\ \cdots \\ \boldsymbol{\Omega} \end{bmatrix} = \boldsymbol{y} \quad (6)$$

где $y = [y_1, y_2, \dots y_T]^T$, — последовательность векторов параметров целевого диктора, $y_j \in R^{1\times p}$, p – размерность вектора параметров, $v=[v_1, v_2, \dots v_Q]^T$ – вектор математических ожиданий для каждой компоненты смеси, где $v_j \in R^{1\times p}$, $\Phi = [\Phi_1, \Phi_2, \dots \Phi_Q]^T$ – матрица регрессионных коэффициентов для всех компонентов смеси при переменной независимой переменной x_j, где $\Phi_j \in R^{p\times p}$, $\Psi = [\Psi_1, \Psi_2, \dots \Psi_Q]^T$ – матрицы регрессионных коэффициентов при переменной независимой переменной y_{i-1}, $\Psi_j \in R^{p\times p}$, $\Omega = [\Omega_1, \Omega_2, \dots \Omega_Q]^T$ – матрицы регрессионных коэффициентов при переменной независимой переменной x_{i+1}, а $\Omega_j \in R^{p\times p}$. Матрицы {P, B, C, D}, размерностью Q½T–1 представляют собой известные характеристики модели и определяются согласно следующим выражениям:

$$\mathbf{P} = \begin{bmatrix} p_1(1) & p_1(2) & \cdots & p_1(T-1) \\ p_2(1) & p_2(2) & \cdots & p_2(T-1) \\ \vdots & \vdots & \ddots & \vdots \\ p_Q(1) & p_Q(2) & \cdots & p_Q(T-1) \end{bmatrix}^T,$$

$$\mathbf{B} = \begin{bmatrix} p_1(1)\Sigma_1^{x^{-1}}(x_1-\mu_1^x) & p_1(2)\Sigma_1^{x^{-1}}(x_2-\mu_1^x) & \cdots & p_1(T-1)\Sigma_1^{x^{-1}}(x_{T-1}-\mu_1^x) \\ p_2(1)\Sigma_2^{x^{-1}}(x_1-\mu_2^x) & p_2(2)\Sigma_2^{x^{-1}}(x_2-\mu_2^x) & \cdots & p_2(T-1)\Sigma_2^{x^{-1}}(x_{T-1}-\mu_2^x) \\ \vdots & \vdots & \ddots & \vdots \\ p_Q(1)\Sigma_Q^{x^{-1}}(x_1-\mu_Q^x) & p_Q(2)\Sigma_Q^{x^{-1}}(x_2-\mu_Q^x) & \cdots & p_Q(T-1)\Sigma_Q^{x^{-1}}(x_{T-1}-\mu_Q^x) \end{bmatrix}^T,$$

$$\mathbf{C} = \begin{bmatrix} p_1(1)\Sigma_1^{y^{-1}}(y_1-\mu_1^y) & p_1(1)\Sigma_1^{y^{-1}}(y_2-\mu_1^y) & \cdots & p_1(T-1)\Sigma_1^{y^{-1}}(y_{T-1}-\mu_1^y) \\ p_2(1)\Sigma_2^{y^{-1}}(y_1-\mu_2^y) & p_2(1)\Sigma_2^{y^{-1}}(y_2-\mu_2^y) & \cdots & p_2(T-1)\Sigma_2^{y^{-1}}(y_{T-1}-\mu_2^y) \\ \vdots & \vdots & \ddots & \vdots \\ p_Q(1)\Sigma_Q^{y^{-1}}(y_1-\mu_Q^y) & p_Q(1)\Sigma_Q^{y^{-1}}(y_2-\mu_Q^y) & \cdots & p_Q(T-1)\Sigma_Q^{y^{-1}}(y_{T-1}-\mu_Q^y) \end{bmatrix}^T,$$

$$\mathbf{D} = \begin{bmatrix} p_1(2)\Sigma_1^{x^{-1}}(x_1-\mu_1^x) & p_1(3)\Sigma_1^{x^{-1}}(x_2-\mu_1^x) & \cdots & p_1(T)\Sigma_1^{x^{-1}}(x_T-\mu_1^x) \\ p_2(2)\Sigma_2^{x^{-1}}(x_1-\mu_2^x) & p_2(3)\Sigma_2^{x^{-1}}(x_2-\mu_2^x) & \cdots & p_2(T)\Sigma_2^{x^{-1}}(x_T-\mu_2^x) \\ \vdots & \vdots & \ddots & \vdots \\ p_Q(2)\Sigma_Q^{x^{-1}}(x_1-\mu_Q^x) & p_Q(3)\Sigma_Q^{x^{-1}}(x_2-\mu_Q^x) & \cdots & p_Q(T)\Sigma_Q^{x^{-1}}(x_T-\mu_Q^x) \end{bmatrix}^T,$$

Легко видеть, что, выполняя замену $A = [P\ B\ C\ D]$, и $\chi = [v\ \Phi\ \Psi\ \Omega]$, уравнение (6) можно привести к нормальной форме. Тогда задача по нахождению неизвестных параметров $\{v, \Phi, \Psi, \Omega\}$ формулируется как задача оптимизации, для решения которой воспользуемся методом наименьших квадратов. Для этого представим выражение (6) в следующем виде:

$$A\cdot\chi = y \Rightarrow A^TA\cdot\chi = yA^T,$$

тогда решение ищется в виде:

$$\chi_{opt} = (A^TA)^{-1}yA^T. \qquad (7)$$

Сложности, которые могут возникнуть при решении данного уравнения, связаны с возможной необходимостью инверсии плохо обусловленных матриц большой размерности, с тенденцией роста количества параметров системы. Общая размерность матрицы, требующей инверсии в правой части выражения (7), зависит от количества компонент смеси Q и размерности векторов параметров p, определяется как $(3½Q½p+p)^2$.

Решение данной проблемы возможно с использованием диагональных ковариационных матриц, вместо их полных версий, а также использованием декомпозиции на основе разложения Холецкого [19]. При практической реализации данного метода использовался алгоритм решения уравнения на основе метода наименьших квадратов с использованием методик, разработанных для решения систем линейных уравнений с плохо обусловленными матрицами коэффициентов большой размерности [20].

2. Гибридная модель конверсии голоса

2.1. Метод спектрального взвешивания

Метод выполнения преобразований, представленный выше, основан на статистических функциях отображения параметров, подходит, в принципе, для всех случаев выполнения операции трансформации над векторами параметров, абстрагировано от их природы. В случае с конверсией голоса это означает, что проблема решается исключительно с математической точки зрения, без учёта специфических характеристик речевого сигнала. В этом, по мнению авторов, заложен некоторый потенциал для совершенствования методов с использованием подходов, более приближённых к физическому смыслу параметров сигнала. Наиболее подходящим для такой задачи видится подход, основанный на спектральном взвешивании, поскольку он ориентирован на выполнение манипуляций с параметрами спектральной огибающей, что как раз и является центральной задачей конверсии голоса.

Основное достоинство метода спектрального взвешивания заключается в том, что он близко связан с акустической теорией речеобразования, в рамках которой еще в работах Фанта [21] было доказано, что формантные частоты различных дикторов связаны нелинейной функцией масштабирования или деформации. Цель метода заключается в нахождении параметров данной функции путем поиска критического пути преобразования между спектральной огибающей исходного и целевого дикторов, относящейся к одному акустическому классу.

Задача формулируется следующим образом. Пусть заданы два спектра $X(f)$ и $Y(f)$ в диапазоне частот $f \in [0; f_{max}]$, оптимальная функция деформации частоты $w(f)$ может быть определена как нелинейная непрерывная функция от f, минимизирующая ошибку, заданную выражением:

$$\varepsilon = \int_0^{f_{max}} (\log|X(f)| - \log|Y(w(f))|)^2 df ,$$

Целью методов конверсии голоса на базе спектрального взвешивания является трансформация частотной оси исходного спектра с использованием специальной функции взвешивания или деформации $w(f)$ так, чтобы сконвертированный спектр максимально соответствовал целевому спектру. Более того, $w(f)$ не должна быть единственной: для различных фонем или групп фонем могут понадобиться различные функции деформации.

2.2. Схема гибридной модели конверсии голоса

В настоящей работе предлагается рассмотреть возможность использования комбинации методов на основе множественных гауссовых смесей и спектраль-

ного взвешивания. За счёт сочетания возможностей статистических методов на основе регрессионного анализа, хорошо справляющихся с задачей описания акустических пространств дикторов, и такой техники обработки сигнала как спектральное взвешивание, тесно связанной с физической природой речевого сигнала, возможно осуществлять трансформацию спектральной огибающей без внесения существенных артефактов в результирующий речевой сигнал, при этом сохранив достаточно высокие характеристики узнаваемости.

Учитывая то, что средние вектора q-й компоненты натренированной модели гауссовых смесей μ_q^x и μ_q^y, информация о положении формант, находящихся на соответствующих участках огибающей данных векторов, может быть использована для определения кусочно-линейной функции деформации частоты $W_q(f)$. Это процесс, изображённый на *рис. 3*, возможен благодаря высокой степени корреляции расположения формант, обнаруживающейся для средних векторов исходного и целевого дикторов, принадлежащих одному акустическому классу. Для модели гауссовых смесей, состоящей из Q компонент, необходимо получить Q различных функций $\{W_q(f)\}$.

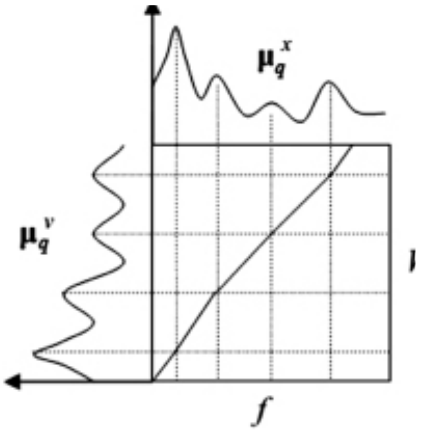

Рис. 3. Функция деформации спектра

Можно предположить, что фонемы с аналогичной формантной структурой, относящейся к одной и той же компоненте смеси, как было отмечено ранее, должны быть связаны с помощью одинаковых функций деформации. С другой стороны, имея параметрическое представление фрейма сигнала x, вероятность принадлежности его к q-й компоненте смеси определяется выражением (3а). Таким образом, центральная идея спектрального взвешивания заключается в вычислении различных функций частотного масштабирования для каждого фрейма входного сигнала, как линейной комбинации Q базисных функций $\{W_q(f)\}$, с использованием апостериорных вероятностей к классам $\{p_q(f_x)\}$ как весов. Выражение, определяющее метод спектрального взвешивания:

$$W(x,f) = \sum p_q(x) W_q(f).$$

Спектр сигнала текущего фрейма должен быть преобразован соответствующей функцией деформации $W_q(f)$, поэтому функция отображения на основе использования метода взвешенной деформации определяется следующим образом:

$$S'_i(f) = G_i(f)\, S_i(W(x_i,f)), \qquad (8)$$

где i – номер фрейма сигнала, $S_i(f)$ и $S'_i(f)$ – исходный и результирующий спектр сигнала соответственно, $W(x_i,f)$ – функция деформации частоты для i-го фрейма сигнала, x_i – вектор параметров спектральной огибающей исходного диктора на i-м фрейме. $G_i(f)$ – фильтр, корректирующий энергию сигнала на i-м фрейме сигнала. Данный фильтр необходим, поскольку представленная процедура изменяет только положение формант на частотной оси, тогда как их мощности и полосы пропускания остаются практически неизменными, что приводит к неправильному распределению энергии по частотам. Манипуляция напрямую этими параметрами в преобразованном спектре может негативно сказаться на характеристиках натуральности восстанов-

ленного сигнала. Однако данная проблема успешно может быть решена путём использования полученной на этапе обучения модели гауссовых смесей, путём конверсии вектора параметров огибающей $F(x)$ согласно выражению (4) и получении сконвертированного спектра $\hat{S}_i'(f)$. То есть данный спектр был бы получен в результате конверсии при отсутствии спектрального взвешивания, но он бы учитывал изменения не только распределения частот, но и энергии сигнала по ним. Поэтому корректирующий фильтр может быть представлен как дискретный набор коэффициентов, сглаживающих усиления $G_i(f) = \hat{S}_i'(f)/S_i'(f)$ в частотной области. Таким образом, производится правильное распределение энергии по частотам сигналов и при этом не наблюдается существенной деградации результирующей огибающей и, как следствие, восстановленного сигнала.

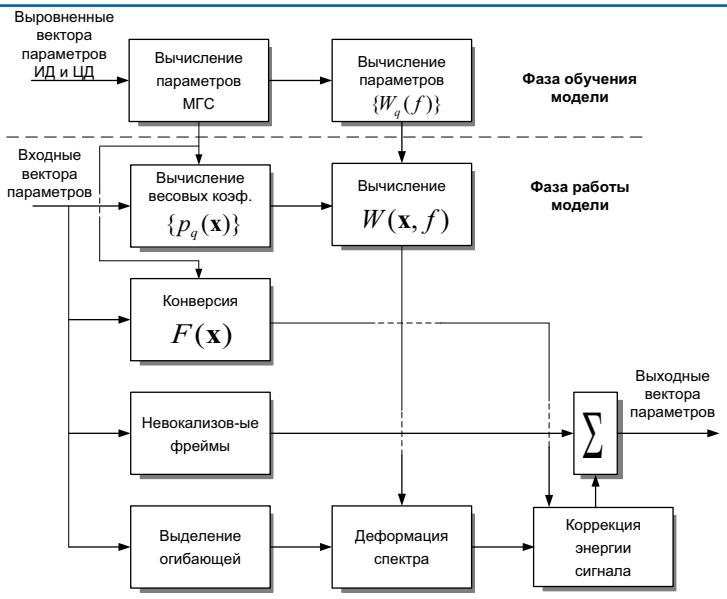

Рис. 4. Схема гибридной модели конверсии голоса на основе множественной регрессионной функции конверсии и методе спектрального взвешивания

С учётом невокализованных фрагментов речи, конверсия которых не производится, схема метода конверсии на основе спектрального взвешивания представлена на *рис. 4*.

В процессе фазы обучения на вход системы поступают выровненные по времени последовательности векторов параметров спектральных огибающих ИД и ЦД, на их основе производится кластеризация акустических пространств дикторов и поиск характеристик моделей множественных гауссовых смесей с использованием EM-алгоритма. Далее на базе средних векторов компонент, согласно выражению (7), ищется функция деформации для каждой из смесей.

Во время фазы работы на вход системы поступает последовательность векторов только ИД, производится расчёт вероятностей принадлежности вектора к q-й компоненте смеси $\{p_q(x)\}$, а также функции деформации для текущего фрейма $W(x,f)$. Далее над спектральной огибающей текущего фрейма, в случае, если он является вокализованным, производится трансформация спектра согласно полученной функции деформации. Невокализованные фреймы

без изменений передаются на выход системы. Затем рассчитываются опорные значения амплитуд с использованием функции отображения $F(x)$ на основе множественной регрессии, по выражению (6), и выполняется расчёт корректирующего фильтра, а также поправок распределения энергии по частотам.

Таким образом, рассмотренный метод спектрального взвешивания позволяет использовать сильные стороны статистических методов на основе моделей множественных гауссовых смесей, для эффективного описания акустического пространства диктора, и в то же время в процессе конверсии использовать возможности метода спектрального взвешивания, имеющего глубинную связь с физической природой речевого сигнала. Это должно поднять характеристики качества восстанавливаемого сигнала, при этом сохранив относительно высокие показатели узнаваемости для сконвертированного сигнала.

3. Экспериментальная часть

Эксперименты проводились на фонетически сбалансированном наборе фраз, включающем по 90 аудиозаписей одинаковых предложений для четырёх дикторов: двух мужчин и двух женщин. В дальнейшем в экспериментах дикторы мужского пола условно обозначены как $m1$ и $m2$, а дикторы женского, как $f1$ и $f2$, соответственно. Средняя длительность одной фразы составляла 5-6 с. Аудиофайлы были закодированы в формате *wav*, с частотой дискретизации 16000кГц и разрядностью сетки квантования в 16 бит. Размер тестовой выборки составлял десять фраз, не входящих в обучающий набор. Анализ и синтез сигнала производился с использованием модели сигнала на базе взвешенной интерполяции спектра, в зарубежной литературе получившей акроним STRAIGHT [22]. После анализа сигнала использовалась параметризация огибающей спектра на основе метода линейного предсказания с использованием представления коэффициентов фильтра в виде вектора линейных спектральных частот 24-го порядка. Для временного масштабирования использовался алгоритм временного выравнивания на основе динамического программирования.

В работе предлагается применить в качестве объективной оценки близости сконвертированного сигнала по отношению к целевому метрику, основанную на кепстральном расстоянии между спектральными огибающими, в шкале Мелов. Данная метрика была выбрана нами, поскольку кепстральные параметры, в отличие от, например, коэффициентов линейного предсказания, обладают наименьшей степенью корреляции между параметрами возбуждения и тембральными характеристиками речевого тракта, заложенными в огибающей спектра. Также важным является факт, что расстояние между ними определяется в психоакустической шкале Мелов, что более обоснованно позволяет интерпретировать значение оценки как степени восприятия искажения восстановленного сигнала человеком [23]. В ходе экспериментов над тестовыми сигналами заново производился их анализ согласно той же модели представления сигнала и с теми же параметрами, что и на этапе конверсии. Однако огибающая спектра была закодирована уже не с помощью коэффициентов линейного предсказания, а при помощи мел-кепстральных коэффициентов 16-го порядка. Оценка рассчитывалась как средняя квадратичная ошибка преобразования согласно выражению:

$$\varepsilon = \frac{1}{N}\sum_{n=1}^{N} \| CC\{\hat{y}_n\} - CC\{y_n\} \|^2,$$

где $CC\{y_{\wedge n}\}$ и $CC_q\{y_n\}$ – мел-кепстральное представление результирующей и исходной спектральной огибающей соответственно. Приведённые в статье методы закодированы следую-

щим образом: МНК – метод на основе функции отображения на базе наименьших квадратов (2); СПР – метод на основе совместной плотности распределения вероятностей (4); МРГ – метод на основе множественной регрессии (5); СПВ – метод на основе спектрального взвешивания (8). Результаты конверсии оценивались для трёх направлений конверсии.

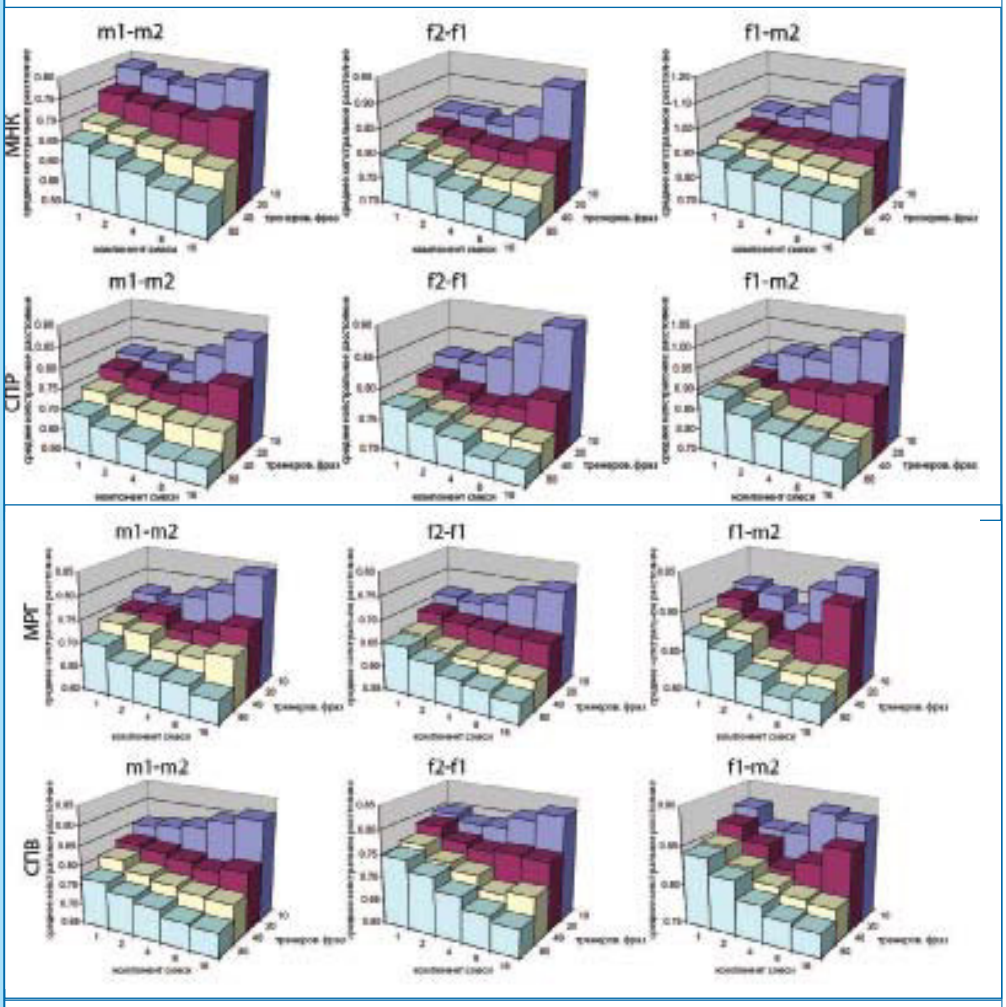

Рис. 5. Результаты тестов для четырёх методов по измерению объективных оценок

Из анализа результатов эксперимента видно, что все методы имеют общий тренд к уменьшению величины среднеквадратической ошибки при увеличении размерности модели множественных гауссовых и количества тренировочных фраз. В особенности данная тенденция хорошо проявляется при увеличении числа компонент модели более 6 и размера обучающей выборки более 40 фраз. При более детальном рассмотрении очевидно, что метод на основе спектрального взвешивания показывает почти во всех случаях самые низкие значения искажений. Что подтверждает ожидаемую эффективность для данного метода. На втором месте следует рассмотренный чуть ранее метод на основе функции конверсии, основанный на множественной регрессионной модели, что также является ожидаемым фактом с точки зрения объективных оценок.

Заключение

В статье были рассмотрены вопросы, связанные с развитием методов и моделей такого направления речевых исследований, как конверсия голоса. Были представлены новые методы на основе множественной регрессии и взвешенной деформации спектра. Первый, имеющий в своей основе статистические методы обработки данных, позволяет с использованием регрессионной функции третьего порядка учесть зависимости более высокого уровня в процессе обучения и работы алгоритма конверсии, нежели стандартные подходы. На этапе обучения это позволяет рассматривать взаимодействие не только между двумя параллельными фреймами, но и учесть присутствие свойств эргодичности в речевом сигнале, рассмотрев эти связи в обучающей последовательности как Марковский процесс. Второй — за счёт сочетания сильных сторон статистических методов на основе регрессионного анализа, хорошо справляющихся с задачей описания акустических пространств дикторов, и такой техники обработки сигнала, как спектральное взвешивание, тесно связанной с физической природой речевого сигнала. Он позволяет осуществлять трансформацию спектральной огибающей без внесения существенных артефактов в результирующий речевой сигнал, при этом сохранив достаточно высокие характеристики узнаваемости. Эффективность рассмотренных методов подтверждается результатами экспериментов, представленными в материалах статьи.

Литература

1. Abe M., Nakamura S., Shikano K. Voice conversion through vector quantization // Proc. of International Conference on Acoustics, Speech and Signal Processing. New York, 1988. P. 655–658.

2. Moulines E., Sagisaka Y. Voice conversion: State of the art and perspectives // Speech Communication. 1995. Vol. 16, K. 125–224.

3. Лобанов Б. М., Цирульник Л. И. Компьютерный синтез и распознавание речи. Минск: Белорусская наука, 2008.

4. Duxans H. Voice conversion applied to text-to-speech systems: Ph.D. thesis // Universitat Politecnica de Catalyunya. Barcelona, 2006. May.

5. Valbret H., Moulines E., Tubach J.P. Voice transformation using PSOLA technique // Proc. of International Conference on Acoustics, Speech and Signal Processing. Vol. 1. 1992. P. 145–148.

6. Arslan L. Speaker transformation algorithm using segmental codebooks (STASC) // Speech Communication. 1999. Vol. 28, no. 3. P. 211–226.

7. Sundermann D., Hoge H., Bonafonte A. Text-independent voice conversion based on unit selection // Proc. of International Conference on Acoustics, Speech and Signal Processing. Vol. 1. 2006.

8. Narendranath M., Murthy H., Rajendran S., Yegnanarayana N. Transformation of formants for voice conversion using artificial neural networks // Speech Communication. 1995. Vol. 16, no. 2. P. 207–216.

9. Machado A. F., Queiroz M. Voice conversion: a critical survey // Open access article. 2010. URL: http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf

10. Азаров И.С., Петровский А.А. Система конверсии голоса в реальном масштабе времени с текстонезависимым обучением на основе гибридного параметрического описания речевых сигналов // Цифровая обработка сигналов. 2012. №2. С. 15–23.

11. Павловец А.С., Лившиц М.З., Личачев Д.С., Петровский А.А. Конверсия голоса с использованием модели сепарации речевого сигнала на компоненты «гармоники + шум» и переходные фреймы // Речевые технологии. 2008. №4. С. 37–50.

12. Анализаторы речевых и звуковых сигналов: методы, алгоритмы и практика (с MATLAB примерами) / под редакцией А.А. Петровского. Минск : Бестпринт, 2009.

13. Stylianou Y., Cappe O., Moulines E. Statistical methods for voice quality transformation // Proc. of European Conference on Speech Communication and Technology. Madrid, 1995. P. 447–450.

14. Bishop C. M. Pattern Recognition and Machine Learning (Information Science and Statistics). Secaucus, NJ, USA : Springer-Verlag New York, Inc., 2007.
15. Kay S. M. Fundamentals of statistical signal processing: estimation theory. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.
16. Stylianou Y., Cappe O., Moulines E. Continuous probabilistic transform for voice conversion // Proc. of International Conference on Acoustics, Speech and Signal Processing. Vol. 6. 1998. P. 131–142.
17. Kain A., Macon M.W. Spectral voice conversion for text-to-speech synthesis // Proc. of International Conference on Acoustics, Speech and Signal Processing. 1998. P. 285–288.
18. Kain A., Macon M.W. Text-to-speech voice adaptation from sparse training data // Proc. of International Conference on Spoken Language Processing. 1998. P. 2847–2850.
19. Вержбицкий В. Основы численных методов. Москва : Высшая школа, 2009.
20. Boyd S. A matlab solver for large-scale-regularised least squares problems. 2012.URL: http://www.stanford.edu/ boyd/l1_Ls (online; accessed: 25.06.2013).
21. Fant G., Kruckenberg A., Nord L. Prosodic and segmental speaker variations // Speech Communication. 1991. Vol. 10, no. 5-6. P. 521–531.
22. Kawahara H., Masanori M. Technic foundations of tandem-straight, a speech analysis, modification and synthesis framework // SADHANA - Academy Proceedings in Engineering Sciences. Vol. 36 of 5. 2011. P. 713–722.
23. Hu Y., Loizou P.C. Evaluation of objective quality measures for speech enhancement // IEEE Transactions on Audio, Speech & Language Processing. 2008. Vol. 16, no. 1, P. 229–238.

Сведения об авторах:

Захарьев Вадим Анатольевич,
аспирант, окончил Белорусский государственный университет информатики и радиоэлектроники, факультет информационных технологий и управления, специальность — «Информационные технологии и управление в технических системах». Область научных интересов: цифровая обработка сигналов, методы распознавания образов и машинного обучения, синтез речи, конверсия голоса. E-mail: zahariev@bsuir.by

Петровский Александр Александрович,
доктор технических наук, профессор, Белорусский государственный университет информатики и радиоэлектроники (бывший Минский радиотехнический институт), кафедра электронных вычислительных средств. Главные научные интересы лежат в области цифровой обработки сигналов речи и звука для целей компрессии, распознавания, редактирования шума, а также проектирование проблемно-ориентированных средств вычислительной техники реального времени для систем мультимедиа. Член НТО РЭС им. А.С.Попова, IEEE, EURASIP, AES.