НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2015 Том 15 № 5 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2015 Vol. 15 No 5 ISSN 2226-1494 http://ntv.ifmo.ru/en
УНИВЕРСИТЕТ ИТМО

**УДК 004. 93**

# ПРИМЕНЕНИЕ МЕТОДА ЧАСТИЧНЫХ НАИМЕНЬШИХ КВАДРАТОВ ДЛЯ ОБРАБОТКИ И МОДЕЛИРОВАНИЯ АУДИОВИЗУАЛЬНОЙ РЕЧИ

**А.Л. Олейник**[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: andrey_oleynik@niuitmo.ru

**Информация о статье**
Поступила в редакцию 08.06.15, принята к печати 16.07.15
doi:10.17586/2226-1494-2015-15-5-886-892
Язык статьи – русский

**Ссылка для цитирования:** Олейник А.Л. Применение метода частичных наименьших квадратов для обработки и моделирования аудиовизуальной речи // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15. № 5. С. 886–892.

**Аннотация**

**Предмет исследования**. Рассмотрена задача реконструкции изображения области рта по речевому сигналу с помощью метода частичных наименьших квадратов. Потребность в решении подобных задач возникает при создании методов обработки аудиовизуальной речи, которая содержит в себе звуковую и визуальную составляющие, называемые модальностями. Конкретные задачи, решаемые с помощью таких методов, включают в себя совместное моделирование голоса и динамики движений губ, синхронизацию аудио- и видеопотоков, распознавание эмоций, обнаружение живости (liveness detection). **Метод**. Для решения поставленной задачи применен метод частичных наименьших квадратов. Метод позволяет выделить из исходных данных компоненты, между которыми существует ковариационная связь, и построить на их основе модель регрессии. Преимуществом такого подхода является возможность решения двух базовых задач: выявления скрытых связей между исходными данными (речевым сигналом и изображением области рта) и аппроксимации одних исходных данных по другим. **Основные результаты**. Экспериментальные исследования по реконструкции изображения области рта по речевому сигналу выполнены на аудиовизуальной речевой базе VidTIMIT. Полученные результаты позволяют сделать вывод о возможности применения метода частичных наименьших квадратов для решения задачи реконструкции. **Практическая значимость**. Результаты проведенного исследования позволяют утверждать, что метод частичных наименьших квадратов может быть успешно применен для решения широкого класса задач обработки аудиовизуальной речи: от синхронизации аудио- и видеопотоков до обнаружения живости.

**Ключевые слова**
обработка аудиовизуальной речи, бимодальные речевые системы, метод частичных наименьших квадратов, ЧНК, методы проекции на подпространства, регрессия.

**Благодарности**
Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01). Автор выражает искреннюю признательность научному руководителю, профессору Г.А. Кухареву, и заведующему кафедрой РИС Ю.Н. Матвееву за критические замечания и советы, которые помогли значительно улучшить качество настоящей статьи.

# APPLICATION OF PARTIAL LEAST SQUARES REGRESSION FOR AUDIO-VISUAL SPEECH PROCESSING AND MODELING

**A.L. Oleinik**[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: andrey_oleynik@niuitmo.ru

**Article info**
Received 08.06.15, accepted 16.07.15
doi:10.17586/2226-1494-2015-15-5-886-892
Article in Russian
**For citation:** Oleinik A.L. Application of Partial Least Squares regression for audio-visual speech processing and modeling. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2015, vol. 15, no. 5, pp. 886–892.

**Abstract**

**Subject of Research**. The paper deals with the problem of lip region image reconstruction from speech signal by means of Partial Least Squares regression. Such problems arise in connection with development of audio-visual speech processing methods. Audio-visual speech consists of acoustic and visual components (called modalities). Applications of audio-visual speech processing methods include joint modeling of voice and lips' movement dynamics, synchronization of audio and video streams, emotion recognition, liveness detection. **Method**. Partial Least Squares regression was applied to solve the posed problem. This method extracts components of initial data with high covariance. These components are used to build

regression model. Advantage of this approach lies in the possibility of achieving two goals: identification of latent interrelations between initial data components (e.g. speech signal and lip region image) and approximation of initial data component as a function of another one. **Main Results**. Experimental research on reconstruction of lip region images from speech signal was carried out on VidTIMIT audio-visual speech database. Results of the experiment showed that Partial Least Squares regression is capable of solving reconstruction problem. **Practical Significance**. Obtained findings give the possibility to assert that Partial Least Squares regression is successfully applicable for solution of vast variety of audio-visual speech processing problems: from synchronization of audio and video streams to liveness detection.

**Keywords**
audio-visual speech processing, bimodal speech systems, Partial Least Squares, PLS, subspace methods, regression.

**Acknowledgements**
The work was done under government financial support for the leading universities of the Russian Federation (grant 074-U01). The author expresses his sincere appreciation to Professor Georgy Kukharev, his scientific adviser, and Yuri Matveev, Head of SIS Department, for their critical remarks and advice that significantly improved the paper.

## Введение

Аудиовизуальная речь включает две составляющие (модальности): звуковую, представленную записью речевого сигнала, и визуальную, которая может включать в себя видеозапись лица или движений губ. Интерес к исследованиям, связанным с обработкой речевой информации в нескольких модальностях, стал стремительно нарастать в последние годы по нескольким причинам. Во-первых, бурное развитие и удешевление средств записи звука и видео позволяет создавать недорогие и компактные устройства для сбора многомодальных данных. Во-вторых, одномодальные системы достигли такого уровня развития, при котором крайне трудно обеспечить повышение качества их работы при помощи новых математических методов, что делает экономически обоснованным использование дополнительных технических средств. В-третьих, в определенных условиях применение одномодальных систем сильно затруднено: например, в шумной обстановке трудно выделить голос интересующего нас диктора, но информация о движениях его губ может сильно в этом помочь. Конкретные задачи, решаемые с помощью этих методов, включают в себя совместное моделирование голоса и динамики движений губ, синхронизацию аудио- и видеопотоков, реконструкцию одной модальности по другой, распознавание эмоций, обнаружение живости (liveness detection [1]).

Несмотря на преимущества бимодальных речевых систем, в процессе их разработки возникает ряд трудностей. Сложность обработки данных о разных модальностях (как, например, о голосе или речи) заключается в их различной природе. Так, например, видео снимается с частотой в несколько десятков кадров в секунду, а частота дискретизации звуковых записей составляет тысячи и десятки тысяч килогерц. Кроме того, механизмы восприятия человеком аудиовизуальных образов достаточно сложны и не до конца изучены. Широко известно явление, называемое эффектом МакГурка [2], когда видеозапись лица, произносящего слоги /ga/, сопровождаемая фонограммой с произношением слогов /ba/, воспринимается человеком как /da/.

Звуковая и визуальная модальности речи состоят из структурных элементов, называемых соответственно фонемами и виземами [3]. При этом связь между ними неоднозначна: одна визема может соответствовать множеству фонем. Это обусловлено, в том числе, и тем фактом, что на видеозаписи значительная часть органов речеобразующей системы человека (например, язык, небная занавеска и голосовые связки) частично или полностью скрыта.

В настоящей работе представлена задача реконструкции изображения области рта по голосу с помощью метода частичных наименьших квадратов (ЧНК). Метод ЧНК относится к классу методов проекции на подпространства, которые предполагают поиск собственного базиса с последующим выбором в нем некоторого количества собственных векторов. Другие методы проекции на подпространства включают в себя метод главных компонент [4], линейный дискриминантный анализ [5] и канонический корреляционный анализ [6]. Они были разработаны достаточно давно, однако применять для обработки голоса и изображений лиц их стали только в последние несколько десятилетий, что обусловлено достижениями в области вычислительной техники. Кроме того, эти методы были обобщены на двумерный случай, что позволяет эффективно применять их для обработки изображений лиц [7].

Метод ЧНК выгодно отличает то, что он позволяет одновременно выявлять скрытые связи между входными данными и аппроксимировать их [8]. Более того, существуют реализации метода ЧНК, позволяющие построить регрессионную модель, описывающую зависимость между входными данными [9]. Последнее свойство особенно важно для решения задачи реконструкции одной модальности по другой.

На данный момент существует несколько работ по применению метода ЧНК для обработки голоса и изображений лиц. Рассматриваются задачи распознавания эмоций [10, 11], чтения по губам и идентификации личности [12], совместной обработки изображений лиц в видимом и инфракрасном спектрах [13]. Решение задачи реконструкции одной модальности по другой на основе метода ЧНК позволит применить его для решения гораздо более широкого класса задач: от синхронизации аудио- и видеопотоков до обнаружения живости.

**Метод частичных наименьших квадратов**

Метод ЧНК позволяет выделить из исходных данных компоненты, между которыми существует ковариационная связь. На основе этих компонент может быть построена модель регрессии. Такой подход позволяет не только существенно снизить вычислительные затраты, но и значительно улучшить точность модели по сравнению с линейной регрессией, построенной с помощью метода наименьших квадратов.

На сегодняшний день известны различные виды и реализации метода ЧНК [14]. В настоящей работе используется регрессия ЧНК, основанная на алгоритме NIPALS (Nonlinear Iterative Partial Least Squares) [9]. При выполнении регрессии ЧНК исходные данные интерпретируются как предикторы (независимые переменные) $\mathbf{x}$ и отклики (зависимые переменные) $\mathbf{y}$. В нашем случае в качестве предикторов выступают голосовые признаки, а в качестве откликов – визуальные признаки (изображения области рта). В рамках метода ЧНК строится модель, позволяющая реконструировать отклики по предикторам.

Пусть исходные данные для построения модели регрессии сведены в матрицы $\mathbf{X}$ и $\mathbf{Y}$. Здесь $\mathbf{X} = (\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_N)^T$ – матрица предикторов размера $N \times D_{\mathbf{X}}$, а $\mathbf{Y} = (\mathbf{y}_1\ \mathbf{y}_2\ \ldots\ \mathbf{y}_N)^T$ – матрица откликов размера $N \times D_{\mathbf{Y}}$, где $N$ – объем выборки, а $D_{\mathbf{X}}$ и $D_{\mathbf{Y}}$ – размерности исходных пространств признаков. Важно заметить, что для корректного построения модели исходные данные $\mathbf{X}$ и $\mathbf{Y}$ должны быть отцентрированы, т.е. векторы средних значений

$$\bar{\mathbf{x}} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\,,\ \bar{\mathbf{y}} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i$$

должны быть нулевыми.

Центрированная модель ЧНК имеет вид [15]:

$$\mathbf{X} = \mathbf{T}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}}^T + \mathbf{E}_{\mathbf{X}} = \mathbf{t}_{\mathbf{X},1}\mathbf{p}_{\mathbf{X},1}^T + \mathbf{t}_{\mathbf{X},2}\mathbf{p}_{\mathbf{X},2}^T + \cdots + \mathbf{t}_{\mathbf{X},p}\mathbf{p}_{\mathbf{X},p}^T + \mathbf{E}_{\mathbf{X}},$$

$$\mathbf{Y} = \mathbf{T}_{\mathbf{Y}}\mathbf{P}_{\mathbf{Y}}^T + \mathbf{E}_{\mathbf{Y}} = \mathbf{t}_{\mathbf{Y},1}\mathbf{p}_{\mathbf{Y},1}^T + \mathbf{t}_{\mathbf{Y},2}\mathbf{p}_{\mathbf{Y},2}^T + \cdots + \mathbf{t}_{\mathbf{Y},p}\mathbf{p}_{\mathbf{Y},p}^T + \mathbf{E}_{\mathbf{Y}}.$$

На рис. 1 показано схематическое представление модели ЧНК.

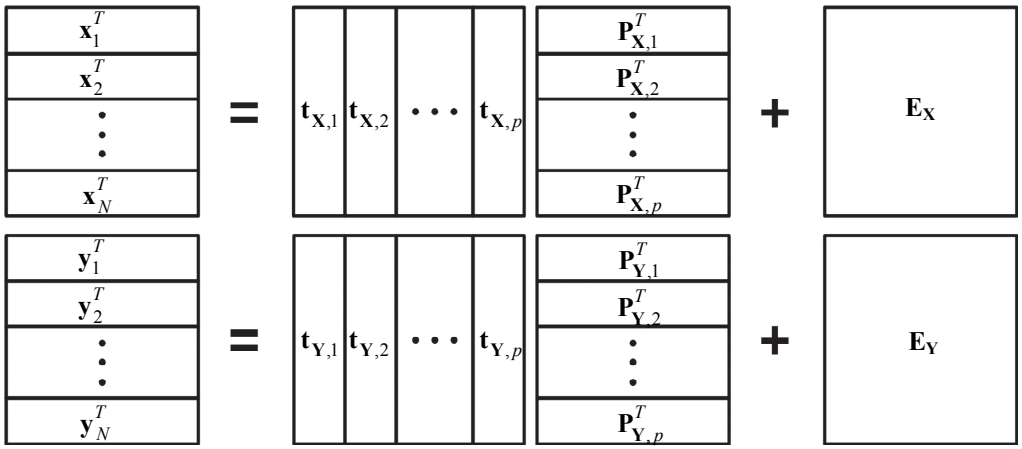


Рис. 1. Модель частичных наименьших квадратов

Здесь использованы следующие обозначения:

$\mathbf{T}_{\mathbf{X}} = (\mathbf{t}_{\mathbf{X},1}\ \mathbf{t}_{\mathbf{X},2}\ \ldots\ \mathbf{t}_{\mathbf{X},p})$ – матрица счетов для предикторов $(N \times p)$;

$\mathbf{T}_{\mathbf{Y}} = (\mathbf{t}_{\mathbf{Y},1}\ \mathbf{t}_{\mathbf{Y},2}\ \ldots\ \mathbf{t}_{\mathbf{Y},p})$ – матрица счетов для откликов $(N \times p)$;

$\mathbf{P}_{\mathbf{X}} = (\mathbf{p}_{\mathbf{X},1}\ \mathbf{p}_{\mathbf{X},2}\ \ldots\ \mathbf{p}_{\mathbf{X},p})$ – матрица нагрузок для предикторов $(D_{\mathbf{X}} \times p)$;

$\mathbf{P}_{\mathbf{Y}} = (\mathbf{p}_{\mathbf{Y},1}\ \mathbf{p}_{\mathbf{Y},2}\ \ldots\ \mathbf{p}_{\mathbf{Y},p})$ – матрица нагрузок для откликов $(D_{\mathbf{Y}} \times p)$;

$\mathbf{E}_{\mathbf{X}}$ – матрица остатков для предикторов $(N \times D_{\mathbf{X}})$;

$\mathbf{E}_{\mathbf{Y}}$ – матрица остатков для откликов $(N \times D_{\mathbf{Y}})$;

$p$ – порядок модели, лежащий в диапазоне $1, \ldots, \operatorname{rang}\mathbf{X}$.

Заметим, что векторы счетов $\mathbf{t}_{\mathbf{X},i}$ и $\mathbf{t}_{\mathbf{Y},i}$ представляют собой проекции исходных данных на направления $\mathbf{w}_{\mathbf{X},i} \in \mathbb{R}^{D_{\mathbf{X}}}$ и $\mathbf{w}_{\mathbf{Y},i} \in \mathbb{R}^{D_{\mathbf{Y}}}$. Вдоль этих направлений достигается локальный максимум ковариации (не корреляции!) между исходными данными. В данной работе для расчета векторов счетов и нагрузок используется алгоритм NIPALS.

**Алгоритм NIPALS** является разновидностью широко известного степенного метода и позволяет рассчитать одну пару векторов счетов $\mathbf{t}_{\mathbf{X}}$ и $\mathbf{t}_{\mathbf{Y}}$, соответствующую направлениям $\mathbf{w}_{\mathbf{X}}$ и $\mathbf{w}_{\mathbf{Y}}$, вдоль которых достигается глобальный максимум ковариации:

$$\{\mathbf{w}_{\mathbf{X}}, \mathbf{w}_{\mathbf{Y}}\} = \operatorname{argmax}_{\mathbf{w}_{\mathbf{X}},\mathbf{w}_{\mathbf{Y}}} \mathbf{w}_{\mathbf{X}}^T\mathbf{X}^T\mathbf{Y}\mathbf{w}_{\mathbf{Y}} = \operatorname{argmax}_{\mathbf{w}_{\mathbf{X}},\mathbf{w}_{\mathbf{Y}}} \mathbf{t}_{\mathbf{X}}^T\mathbf{t}_{\mathbf{Y}}\,, \qquad \textit{где } \|\mathbf{w}_{\mathbf{X}}\|^2 = \|\mathbf{w}_{\mathbf{Y}}\|^2 = 1.$$

Индекс $i$ здесь опущен, так как речь идет только об одной паре векторов счетов.

Можно показать [9], что $\mathbf{w_X}$ и $\mathbf{w_Y}$ могут быть получены как решение задачи на собственные векторы, соответствующее максимальному собственному числу λ:

$$\begin{cases}\mathbf{X}^T\mathbf{Y}\mathbf{w_Y} = \lambda\mathbf{w_X},\\ \mathbf{Y}^T\mathbf{X}\mathbf{w_X} = \lambda\mathbf{w_Y}.\end{cases}$$

Получить искомое решение можно с помощью алгоритма NIPALS [9]. Результат его работы будем записывать в следующей форме:

$$\{\mathbf{t_X}, \mathbf{w_X}, \mathbf{t_Y}, \mathbf{w_Y}, \lambda\} = \text{NIPALS}(\mathbf{X}, \mathbf{Y}).$$

**Построение модели ЧНК.** Применение метода ЧНК для построения модели регрессии предполагает существование линейного отображения, связывающего матрицы счетов [14]:

$$\mathbf{T_Y} = \mathbf{T_X}\mathbf{D} + \mathbf{H}.$$

Здесь $\mathbf{D}$ – матрица размера $p \times p$; $\mathbf{H}$ – матрица остатков размера $N \times p$. Тогда

$$\mathbf{Y} = \mathbf{T_Y}\mathbf{P_Y}^T + \mathbf{E_Y} = (\mathbf{T_X}\mathbf{D} + \mathbf{H})\mathbf{P_Y}^T + \mathbf{E_Y} = \mathbf{T_X}\mathbf{D}\mathbf{P_Y}^T + (\mathbf{H}\mathbf{P_Y}^T + \mathbf{E_Y}) = \mathbf{T_X}\mathbf{Q_Y}^T + \mathbf{F_Y}.$$

Здесь $\mathbf{Q_Y} = \mathbf{P_Y}\mathbf{D}^T$ – матрица размера $D_\mathbf{Y} \times p$; $\mathbf{F_Y}$ – матрица остатков размера $N \times D_\mathbf{Y}$. Таким образом, исходная модель ЧНК запишется в следующей форме:

$$\mathbf{X} = \mathbf{T_X}\mathbf{P_X}^T + \mathbf{E_X} = \mathbf{t}_{\mathbf{X},1}\mathbf{p}_{\mathbf{X},1}^T + \mathbf{t}_{\mathbf{X},2}\mathbf{p}_{\mathbf{X},2}^T + \cdots + \mathbf{t}_{\mathbf{X},p}\mathbf{p}_{\mathbf{X},p}^T + \mathbf{E_X},$$

$$\mathbf{Y} = \mathbf{T_X}\mathbf{Q_Y}^T + \mathbf{F_Y} = \mathbf{t}_{\mathbf{X},1}\mathbf{q}_{\mathbf{Y},1}^T + \mathbf{t}_{\mathbf{X},2}\mathbf{q}_{\mathbf{Y},2}^T + \cdots + \mathbf{t}_{\mathbf{X},p}\mathbf{q}_{\mathbf{Y},p}^T + \mathbf{F_Y}.$$

Заметим теперь, что матрицы исходных данных $\mathbf{X}$ и $\mathbf{Y}$ выражены через общую матрицу счетов $\mathbf{T_X}$. Для построения модели ЧНК в такой форме используется следующий алгоритм [9]:

**Шаг 0.** $\mathbf{X}^0 \leftarrow \mathbf{X}$, $\mathbf{Y}^0 \leftarrow \mathbf{Y}$, $i \leftarrow 1$;

**Шаг 1.** $\{\mathbf{t}_{\mathbf{X},i}, \mathbf{w}_{\mathbf{X},i}, \mathbf{t}_{\mathbf{Y},i}, \mathbf{w}_{\mathbf{Y},i}, \lambda_i\} = \text{NIPALS}(\mathbf{X}^{i-1}, \mathbf{Y}^{i-1})$;

**Шаг 2.** $\mathbf{p}_{\mathbf{X},i} \leftarrow {\mathbf{X}^{i-1}}^T \frac{\mathbf{t}_{\mathbf{X},i}}{\mathbf{t}_{\mathbf{X},i}^T\mathbf{t}_{\mathbf{X},i}}$, $\quad \mathbf{q}_{\mathbf{Y},i} \leftarrow {\mathbf{Y}^{i-1}}^T \frac{\mathbf{t}_{\mathbf{X},i}}{\mathbf{t}_{\mathbf{X},i}^T\mathbf{t}_{\mathbf{X},i}}$;

**Шаг 3.** $\mathbf{X}^i \leftarrow \mathbf{X}^{i-1} - \mathbf{t}_{\mathbf{X},i}\mathbf{p}_{\mathbf{X},i}^T$, $\quad \mathbf{Y}^i \leftarrow \mathbf{Y}^{i-1} - \mathbf{t}_{\mathbf{X},i}\mathbf{q}_{\mathbf{Y},i}^T$;

**Шаг 4.** $i \leftarrow i + 1$;

**Шаг 5.** Повторять шаги 1–4, пока $i \le p$;

**Шаг 6.** Из полученных векторов сформировать матрицы:

- $\mathbf{T_X} = (\mathbf{t}_{\mathbf{X},1}\ \mathbf{t}_{\mathbf{X},2}\ \ldots\ \mathbf{t}_{\mathbf{X},p})$ и $\mathbf{T_Y} = (\mathbf{t}_{\mathbf{Y},1}\ \mathbf{t}_{\mathbf{Y},2}\ \ldots\ \mathbf{t}_{\mathbf{Y},p})$;
- $\mathbf{P_X} = (\mathbf{p}_{\mathbf{X},1}\ \mathbf{p}_{\mathbf{X},2}\ \ldots\ \mathbf{p}_{\mathbf{X},p})$ и $\mathbf{Q_Y} = (\mathbf{q}_{\mathbf{Y},1}\ \mathbf{q}_{\mathbf{Y},2}\ \ldots\ \mathbf{q}_{\mathbf{Y},p})$;
- $\mathbf{E_X} = \mathbf{X}^p$ и $\mathbf{F_Y} = \mathbf{Y}^p$;
- $\mathbf{W_X} = (\mathbf{w}_{\mathbf{X},1}\ \mathbf{w}_{\mathbf{X},2}\ \ldots\ \mathbf{w}_{\mathbf{X},p})$ – матрица весов размера $D_\mathbf{X} \times p$.

Таким образом, приведенный алгоритм позволяет построить модель ЧНК. Отметим, что операция, выполняемая на шаге 3, снижает ранг матрицы $\mathbf{X}$ на единицу. В качестве критерия останова на шаге 5 вместо использования фиксированного $p$ можно, например, оценивать остаточную дисперсию матрицы $\mathbf{X}^{i-1}$ [15]. Матрицы $\mathbf{T_Y}$ и $\mathbf{W_X}$ будут использованы ниже для построения модели регрессии.

**Построение модели регрессии** выполняется в два этапа:

1. проекция исходных данных $\mathbf{X}$ и $\mathbf{Y}$ на подпространства размерности $p$, порожденные базисами $\{\mathbf{w}_{\mathbf{X},i}\}$ и $\{\mathbf{w}_{\mathbf{Y},i}\}$;
2. построение модели линейной регрессии для полученных проекций с помощью метода наименьших квадратов (Ordinary Least Squares, OLS).

Заметим, что, так как операция проекции сводится к умножению на матрицу, построенная модель регрессии может быть записана в исходных системах координат [9]:

$$\mathbf{Y} = \mathbf{X}\mathbf{R} + \mathbf{G}.$$

Здесь $\mathbf{R}$ – матрица регрессии размера $D_\mathbf{X} \times D_\mathbf{Y}$; $\mathbf{G}$ – матрица остатков размера $N \times D_\mathbf{Y}$. Полученные на этапе построения модели ЧНК матрицы весов, нагрузок и счетов позволяют записать выражение для матрицы регрессии $\mathbf{R}$ [9]:

$$\mathbf{R} = \mathbf{W_X}(\mathbf{P_X}^T\mathbf{W_X})^{-1}\mathbf{Q_Y}^T = \mathbf{X}^T\mathbf{T_Y}(\mathbf{T_X}^T\mathbf{X}\mathbf{X}^T\mathbf{T_Y})^{-1}\mathbf{T_X}^T\mathbf{Y}.$$

Таким образом, реконструкция откликов по предикторам сводится к умножению на матрицу регрессии:

$$\mathbf{y} \approx \hat{\mathbf{y}} = \mathbf{R}^T\mathbf{x}.$$

## Экспериментальные исследования

Исследования проводились на аудиовизуальной речевой базе VidTIMIT [16], которая включает в себя аудио- и видеозаписи 42 людей (дикторов), произносящих различные предложения на английском языке. Для проведения исследований использовались язык программирования Python, математический пакет MATLAB, библиотека компьютерного зрения OpenCV, а также библиотека для извлечения речевых признаков [17].

Метод ЧНК применен для реконструкции изображения области рта по речевому сигналу. Как было сказано выше, в качестве предикторов $\mathbf{x}$ выступают голосовые признаки, а в качестве откликов $\mathbf{y}$ – визуальные признаки. На рис. 2 показан процесс выделения голосовых и визуальных признаков. В качестве голосовых признаков используются мел-частотные кепстральные коэффициенты (МЧКК), широко при-

меняемые для решения задач голосовой биометрии и распознавания речи [18]. Векторы МЧКК строятся на основе перекрывающихся сегментов речевого сигнала длительностью 25 мс со сдвигом в 10 мс.

Визуальные признаки выделяются из кадров видеопотока. Прежде всего из каждого кадра извлекается изображение области рта. Далее эти изображения перемасштабируются, что позволяет снизить размерность формируемых векторов визуальных признаков с нескольких тысяч до сотен. Затем уменьшенные изображения преобразуются в векторы посредством конкатенации столбцов. Интерполяция требуется из-за того, что количество векторов голосовых признаков (100 векторов на одну секунду) превосходит количество кадров видеопоследовательности (25 кадров в секунду).

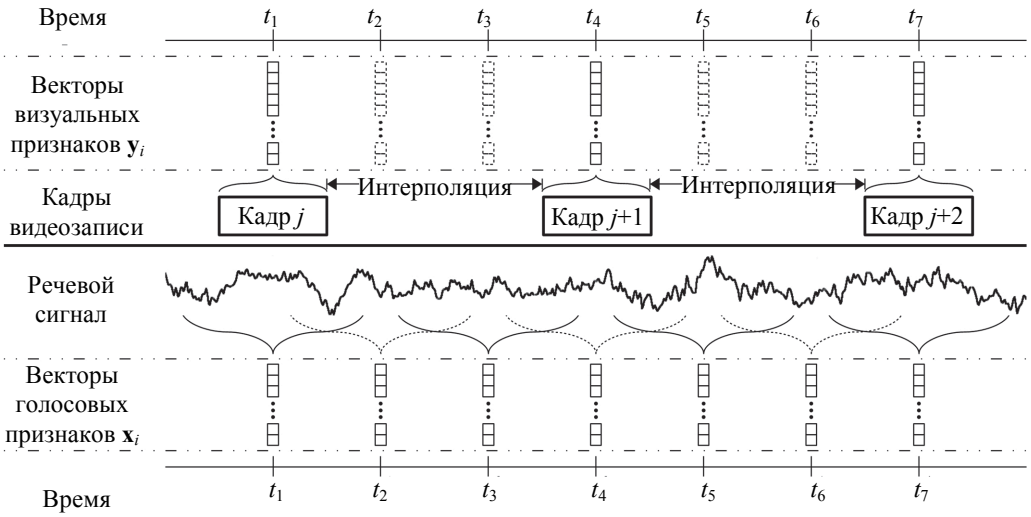


Рис. 2. Выделение визуальных и голосовых признаков

На рис. 3 показана схема проведенного эксперимента. Из голосовых и визуальных признаков формируется обучающая выборка $\mathbf{X}^{train} = (\mathbf{x}_1^{train}\ \mathbf{x}_2^{train}\ ...\ \mathbf{x}_N^{train})^T$, $\mathbf{Y}^{train} = (\mathbf{y}_1^{train}\ \mathbf{y}_2^{train}\ ...\ \mathbf{y}_N^{train})^T$ и тестовая выборка $\mathbf{X}^{test} = (\mathbf{x}_1^{test}\ \mathbf{x}_2^{test}\ ...\ \mathbf{x}_M^{test})^T$, $\mathbf{Y}^{test} = (\mathbf{y}_1^{test}\ \mathbf{y}_2^{test}\ ...\ \mathbf{y}_M^{test})^T$. Обучающая и тестовая выборки не пересекаются. С помощью описанной выше регрессии ЧНК из обучающей выборки формируются параметры регрессии: матрица регрессии $\mathbf{R}$ и векторы средних значений $\bar{\mathbf{x}}$ и $\bar{\mathbf{y}}$. Затем параметры регрессии используются для реконструкции визуальных признаков $\hat{\mathbf{y}}_i^{test}$ по голосовым признакам из тестовой выборки $\mathbf{x}_i^{test}$. Также формируются векторы ошибки реконструкции $\mathbf{e}_i$ (рис. 3). Преобразование векторов $\mathbf{y}_i^{test}$, $\hat{\mathbf{y}}_i^{test}$ и $\mathbf{e}_i$ в изображения является обратным по отношению к конкатенации столбцов, выполняемой на этапе выделения признаков.

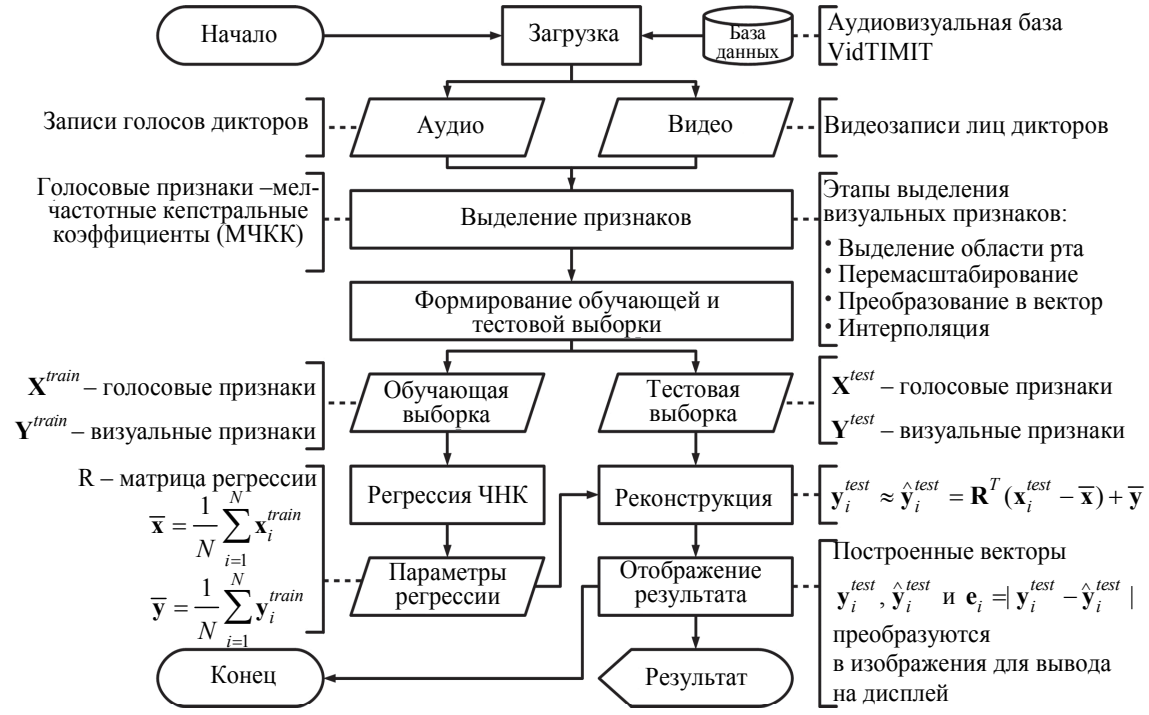


Рис. 3. Схема эксперимента

На рис. 4 представлены исходные и реконструированные по речевому сигналу изображения области рта, а также ошибка реконструкции. Можно отметить наличие сходства между изображениями области рта на рис. 4, а, и рис. 4, б. При этом ошибка реконструкции невелика. С другой стороны, разрешение полученных изображений достаточно мало, что связано с перемасштабированием на этапе выделения визуальных признаков. Кроме того, реконструированные изображения несколько размыты.

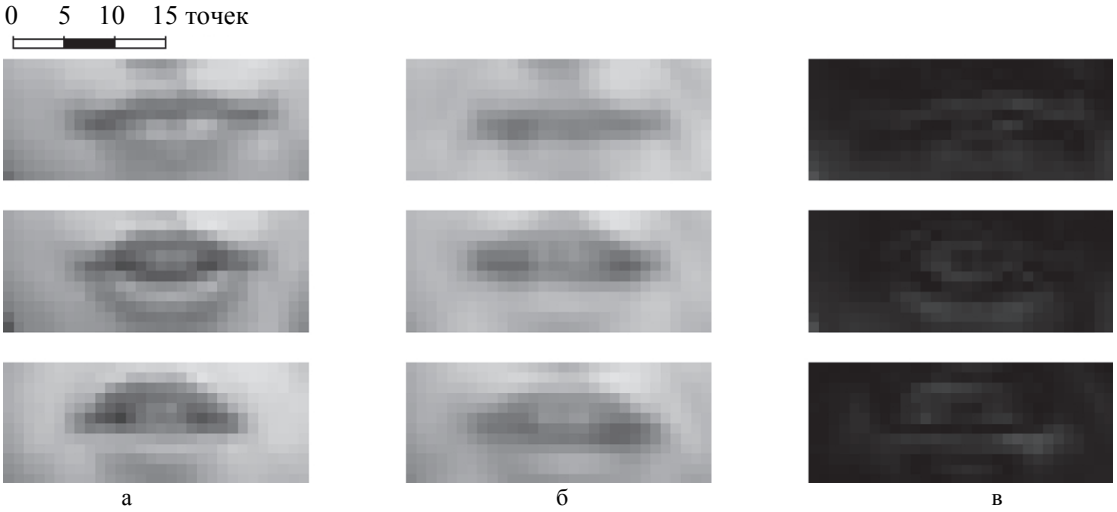


Рис. 4. Результаты экспериментальных исследований: исходные изображения области рта (а), реконструированные по речевому сигналу изображения области рта (б), и ошибка реконструкции (в)

**Заключение**

В работе представлена задача реконструкции изображения области рта по голосу с помощью метода ЧНК. Рассмотрен математический аппарат, лежащий в основе работы, проведены экспериментальные исследования, дана оценка полученным результатам. На их основании можно сделать вывод о применимости метода ЧНК для решения широкого класса задач обработки аудиовизуальной речи.

Можно выделить несколько путей преодоления недостатков предложенного подхода. Во-первых, использование методов предобработки (например, фильтрации и нормализации изображений области рта) может повысить качество обучающей выборки и, следовательно, снизить ошибку реконструкции. Во-вторых, обойти проблему высокой размерности визуальных признаков можно с помощью методов двумерной проекции, описанных в [19]. С их помощью можно добиться лучшего согласования размерностей исходных данных и снизить вычислительные затраты. И наконец, имеет смысл сравнить предложенный подход с решением на основе канонического корреляционного анализа.

**Литература**

1. Chetty G., Wagner M. Liveness detection using cross-modal correlations in face-voice person authentication // Proc. 9th European Conference on Speech Communication and Technology. Lisbon, Portugal, 2005. P. 2181–2184.
2. McGurk H., MacDonald J. Hearing lips and seeing voices // Nature. 1976. V. 264. N 5588. P. 746–748.
3. Aghaahmadi M., Dehshibi M.M., Bastanfard A., Fazlali M. Clustering Persian viseme using phoneme subspace for developing visual speech application // Multimedia Tools and Applications. 2013. V. 65. N 3. P. 521–541. doi: 10.1007/s11042-012-1128-7
4. Pearson K. On lines and planes of closest fit to system of points in space // Philosophical Magazine. 1901. V. 2. N 6. P. 559–572.
5. Fisher R.A. The use of multiple measurements in taxonomic problems // Annals of Eugenics. 1936. V. 7. N 2. P. 179–188.
6. Hotelling H. Relations between two sets of variates // Biometrika. 1936. V. 28. N ¾. P. 321–377.
7. Kukharev G., Kamenskaya E. Application of two-dimensional canonical correlation analysis for face image processing and recognition // Pattern Recognition and Image Analysis. 2010. V. 20. N 2. P. 210–219. doi: 10.1134/S1054661810020136
8. Кухарев Г.А., Каменская Е.И., Матвеев Ю.Н., Щеголева Н.Л. Методы обработки и распознавания изображений лиц в задачах биометрии / под ред. М.В. Хитрова. СПб.: Политехника, 2013. 388 с.
9. De Bie T., Cristianini N., Rosipal R. Eigenproblems in pattern recognition / In: Handbook of Geometric Computing. Ed. E.B. Corrochano. Berlin, Springer, 2005. P. 129–167. doi: 10.1007/3-540-28247-5_5

10. Meng H., Huang D., Wang H., Yang H., Al-Shuraifi M., Wang Y. Depression recognition based on dynamic facial and vocal expression features using partial least square regression // Proc. 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC 2013). Barselona, Spain, 2013. P. 21–29. doi: 10.1145/2512530.2512532
11. Liu M., Wang R., Huang Z., Shan S., Chen X. Partial least squares regression on grassmannian manifold for emotion recognition // Proc. 15th ACM on International Conference on Multimodal Interaction (ICMI'13). Sydney, Australia, 2013. P. 525–530. doi: 10.1145/2522848.2531738
12. Bakry A., Elgammal A. MKPLS: Manifold kernel partial least squares for lipreading and speaker identification // Proc. 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013). Portland, USA, 2013. P. 684–691. doi: 10.1109/CVPR.2013.94
13. Xie Z. Partial least squares regression on DCT domain for infrared face recognition // Proceedings of SPIE – Progress in Biomedical Optics and Imaging. 2014. V. 9230. Art. 92301I. doi: 10.1117/12.2068214
14. Abdi H. Partial least squares regression and projection on latent structure regression (PLS Regression) // Wiley Interdisciplinary Reviews: Computational Statistics. 2010. V. 2. N 1. P. 97–106. doi: 10.1002/wics.51
15. Эсбенсен К. Анализ многомерных данных. Черноголовка: ИПХФ РАН, 2005. 160 с.
16. Sanderson C., Lovell B.C. Multi-region probabilistic histograms for robust and scalable identity inference // Lecture Notes in Computer Science. 2009. V. 5558 LNCS. P. 199–208. doi: 10.1007/978-3-642-01793-3_21
17. Wojcicki K. Mel Frequency Cepstral Coefficient Feature Extraction [Электронный ресурс]. Режим доступа: www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab свободный. Яз. англ. (дата обращения: 2015.06.12).
18. Huang X., Acero A., Hon H.W. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall, 2001. 1008 p.
19. Kukharev G., Tujaka A., Forczmanski P. Face recognition using two-dimensional CCA and PLS // International Journal of Biometrics. 2011. V. 3. N 4. P. 300–321. doi: 10.1504/IJBM.2011.042814

***Олейник Андрей Леонидович*** – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, andrey_oleynik@niuitmo.ru

***Andrei L. Oleinik*** – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, andrey_oleynik@niuitmo.ru