НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2018 Том 18 № 2 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2018 Vol. 18 No 2 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 004.93

ОБРАБОТКА И АНАЛИЗ ЗВУКОВОЙ И ВИЗУАЛЬНОЙ СОСТАВЛЯЮЩИХ РЕЧИ НА ОСНОВЕ ПРОЕКЦИОННЫХ МЕТОДОВ

А.Л. Олейник[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: aoleinik@corp.ifmo.ru

Информация о статье
Поступила в редакцию 23.01.18, принята к печати 25.02.18
doi: 10.17586/2226-1494-2018-18-2-243-254
Язык статьи – русский

Ссылка для цитирования: Олейник А.Л. Обработка и анализ звуковой и визуальной составляющих речи на основе проекционных методов // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 2. С. 243–254. doi: 10.17586/2226-1494-2018-18-2-243-254

Аннотация

Предмет исследования. Рассмотрена задача взаимной реконструкции (преобразования) звуковой и визуальной составляющих (модальностей) речевого сигнала. Аудиозапись голоса представляет звуковую составляющую, а снятая параллельно с ней видеозапись лица человека образует визуальную составляющую. Так как эти модальности обладают различной физической природой, их совместный анализ и обработка сопровождаются рядом трудностей и проблем. Многие из них можно преодолеть с помощью методов взаимной реконструкции. **Метод**. Предложенный подход основан на анализе главных компонент (Principal Component Analysis, PCA), множественной линейной регрессии, регрессии частичных наименьших квадратов (Partial Least Squares, PLS), а также на алгоритме кластеризации K-средних. Также подробно рассмотрены вопросы предобработки исходных данных. В качестве звуковых признаков использованы мел-частотные кепстральные коэффициенты (Mel-Frequency Cepstral Coefficients, MFCC), а в качестве визуальных – набор из 20 опорных точек, представляющих контур рта. **Основные результаты**. В рамках экспериментальных исследований выполнена реконструкция опорных точек контура рта из MFCC. Эксперименты проведены на аудиовизуальной англоязычной базе VidTIMIT. Представлены варианты реализации предложенного подхода на основе PCA и регрессии PLS с кластеризацией и без нее (четыре варианта). Количественная (объективная) и качественная (субъективная) оценки подтвердили работоспособность предложенного подхода; наилучшие результаты показала реализация на основе регрессии PLS с предварительной кластеризацией. **Практическая значимость**. На основе предложенного подхода могут быть разработаны бимодальные биометрические системы, управляемые голосом виртуальные двойники («аватары»), системы контроля доступа к мобильным устройствам и другие решения в области аудиовизуальных человеко-машинных интерфейсов. Показано, что при правильной организации вычислений использование методов PCA и PLS позволяет значительно сократить вычислительные затраты. Отказ от кластеризации также позволяет повысить быстродействие за счет некоторого снижения качества реконструкции.

Ключевые слова
бимодальные речевые системы, реконструкция, анализ главных компонент, кластеризация, метод частичных наименьших квадратов, регрессия

Благодарности
Исследования выполнены за счет стартового финансирования Университета ИТМО.

AUDIO-VISUAL SPEECH PROCESSING AND ANALYSIS BASED ON SUBSPACE PROJECTIONS

A.L. Oleinik[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: aoleinik@corp.ifmo.ru

Article info
Received 23.01.18, accepted 25.02.18
doi: 10.17586/2226-1494-2018-18-2-243-254
Article in Russian
For citation: Oleinik A.L. Audio-visual speech processing and analysis based on subspace projections. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 2, pp. 243–254 (in Russian). doi: 10.17586/2226-1494-2018-18-2-243-254

Abstract

Subject of Research. The paper deals with the problems of the mutual reconstruction (transformation) of acoustic and visual components (modalities) of speech. Audio recording of voice represents the acoustic component whereas the parallel video recording of the speaker's face comprises the visual component. Because of the different physical nature of these modalities, their mutual analysis is accompanied by numerous difficulties. Reconstruction methods can be used to overcome these difficulties. **Method.** The proposed approach is based on Principal Component Analysis (PCA), Multiple Linear Regression (MLR), Partial Least Squares regression (PLS regression) and K-means clustering algorithm. Moreover, attention is paid to data preprocessing. Mel-frequency cepstral coefficients (MFCCs) are used as acoustic features, and twenty key points, which represent the mouth contour, comprise visual features. **Main Results**. The experiments on the reconstruction of the mouth contour from the MFCCs are presented. The experiments were carried out on VidTIMIT dataset of audio-visual phrase recordings in English. Four variants of the proposed approach were tested and evaluated. They are based on PCA and PLS regression with clustering and without it. Quantitative (objective) and qualitative (subjective) assessment confirmed the efficiency of the proposed approach. The implementation based on PLS regression with preliminary clustering led to the best results. **Practical Relevance**. The proposed approach can be used to develop various bimodal biometric systems, voice-driven virtual "avatars", mobile access control systems and other useful human-computer interaction solutions. Moreover, it is shown that, given the proper implementation, PCA and PLS reduce significantly the computational complexity of the reconstruction operation. In addition, the clustering step can be omitted to increase additionally the processing speed at the cost of slightly lower reconstruction quality.

Keywords

bimodal speech systems, reconstruction, principal component analysis, clustering, partial least squares, regression

Acknowledgements

This work was partially supported by ITMO University start-up funding.

Введение

Область применения аудиовизуальных речевых систем весьма широка и включает в себя робототехнику, мобильные приложения и человеко-машинные интерфейсы. Постоянное удешевление и совершенствование аудио- и видеозаписывающих устройств способствует повсеместному внедрению таких систем. Эти факторы порождают потребность в разработке методов совместной обработки и анализа звуковой и визуальной составляющих речи (называемых модальностями).

Существенные различия в структуре речевого сигнала и видеозаписи движений лица порождают ряд проблем, возникающих при их совместной обработке. Одна из таких проблем – разная структура аудио- и видеозаписей. Аудиозапись представляет собой последовательность дискретных отсчетов, следующих с частотой в тысячи и десятки тысяч килогерц, в то время как видеозапись снимается с частотой в несколько десятков кадров (т.е. изображений) в секунду. Приведение таких данных к единой временной шкале в некоторых случаях вызывает определенные затруднения. Эту проблему можно частично решить с помощью высокоскоростных камер [1], однако их использование возможно далеко не всегда.

Кроме того, существует явление, называемое эффектом МакГурка [2]. Он проявляется, например, в том, что видеозапись лица, произносящего слоги /ga/, сопровождаемая фонограммой с произношением слогов /ba/, воспринимается человеком как /da/.

Еще одна трудность заключается в том, что соответствие между структурными элементами звуковой и визуальной составляющих речи – фонемами и виземами – не является взаимно однозначным. Это обусловлено тем, что большая часть артикуляторных органов человека скрыта от стороннего наблюдателя и не отображена на видеозаписи.

Для объединения звуковых и визуальных данных применяют процедуру под названием «многомодальное слияние» (далее – «слияние») [3]. Методы слияния делятся на два основных вида: слияние на уровне решений (позднее слияние) и слияние на уровне признаков (раннее слияние).

Слияние на уровне решений может рассматриваться как расширение существующих одномодальных подходов. В этом случае звуковая и визуальная составляющие обрабатываются отдельно (например, для распознавания личности используются две подсистемы – лицевая и голосовая), после чего результаты работы базовых алгоритмов объединяются. Операция слияния реализуется при помощи относительно простых подходов, таких как голосование, линейное взвешивание или логические операторы [3]. Главными преимуществами слияния на уровне решений являются простота и возможность использования готовых (одномодальных) алгоритмов обработки и выделения признаков. С другой стороны, такой подход никак не учитывает информацию о скрытых связях между модальностями.

Этого недостатка лишены подходы к слиянию на уровне признаков, предполагающие формирование общего компактного признакового представления, к которому затем применяют стандартные алгоритмы машинного обучения. Методы, реализующие данный подход, весьма разнообразны и основаны на сдвоенных скрытых марковских моделях [4, 5], методе опорных векторов [6] и самоорганизующихся картах Кохонена [7].

Важно отметить широкое применение методов глубокого обучения. В частности, для обработки и генерации последовательных данных используют рекуррентные нейронные сети, а также такие методы, как Long Short-Term Memory (LSTM) [8] и Echo State Network (ESN) [9]. Для обработки изображений в

настоящее время чаще всего используют сверточные нейронные сети [10]. В аудиовизуальных системах обработки речи эти подходы нередко комбинируются. С помощью методов глубокого обучения решаются задачи удаления шума на основе визуальной информации [11], распознавания речи [12], поиска диктора, произносящего заданную фразу, на кадрах видеозаписи [13].

Несмотря на гибкость и универсальность, методы глубокого обучения не лишены недостатков, к которым относятся сложность, трудность интерпретации полученных моделей, склонность к переобучению, большие вычислительные затраты, отсутствие методологии разработки архитектур и обучения, а также потребность в объемной обучающей выборке.

С учетом того, что существующие аудиовизуальные базы зачастую не обладают таким объемом, как, например, базы изображений лиц, последний фактор является существенным препятствием для применения методов глубокого обучения.

Альтернативой являются проекционные методы – анализ главных компонент (Principal Component Analysis, PCA), метод частичных наименьших квадратов (Partial Least Squares, PLS), канонический корреляционный анализ (Canonical Correlation Analysis, CCA) и линейный дискриминантный анализ (Linear Discriminant Analysis, LDA). К достоинствам проекционных методов относятся универсальность, простота, интерпретируемость, математическая обоснованность, отсутствие необходимости в объемной обучающей выборке и относительно небольшие вычислительные затраты. Кроме того, для обработки изображений были разработаны двумерные модификации этих методов [14]. С помощью проекционных методов решаются задачи распознавания эмоций [15, 16], чтения по губам [17], синхронизации [18], разделения источников сигнала [19] и аудиовизуального распознавания речи [20].

В настоящей работе предложен метод выявления скрытых связей между модальностями и их взаимной реконструкции. Возможные приложения методов реконструкции включают бимодальный синтез речи, создание виртуальных аватаров (двойников), антиспуфинг (выявление попыток обхода биометрической защиты).

Предложенный метод основан на проекционных и регрессионных методах с применением алгоритмов кластеризации. За основу взяты методы PCA и PLS. Для решения задачи реконструкции наилучшим образом подходит именно PLS, так как он предполагает не только поиск новых «скрытых» переменных, но и решение регрессионной задачи [21, 22]. В случае PCA для построения регрессионной модели используется метод множественной линейной регрессии. Кластеризация на основе алгоритма K-средних используется как средство повышения точности модели реконструкции: в этом случае линейная модель заменяется кусочно-линейной.

Помимо собственно реконструкции, в работе рассмотрены вопросы предобработки и подготовки звуковых и визуальных данных. Несмотря на то, что эти задачи носят вспомогательный характер, правильная подготовка исходных данных нередко вносит в конечный результат не меньший вклад, чем собственно алгоритмы анализа и обработки данных.

Предобработка и подготовка исходных данных

Используемые алгоритмы машинного обучения применяются не к исходным звуковым и визуальным данным, а к выделенным из них признакам, сведенным в матрицы **X** и **Y** соответственно. Эти матрицы содержат по N строк, где N – количество векторов признаков.

Исходные аудиозаписи представляют собой векторы дискретных отсчетов с заданной частотой дискретизации, а видеозаписи движений лиц – последовательности кадров (цветных изображений). Для преобразования этих данных в матрицы **X** и **Y** выполняется последовательность операций по их подготовке и выделению признаков, представленная в табл. 1.

Звуковые признаки	Визуальные признаки
• Удаление шума • Удаление тишины в начале и конце каждой записи • Выделение признаков: 13 MFCC + logE • Оценка производных: Δ и ΔΔ • Z-нормализация (отдельно для каждой фразы)	• Детектирование лица • Детектирование опорных точек лица • Выделение опорных точек контура рта (20 точек) • Нормировка опорных точек контура рта: компенсация поворота, масштаба и размера рта • Z-нормализация (отдельно для каждой фразы)
• Синхронизация, интерполяция с добавлением постоянного сдвига	
Выход: матрица размера N×42	**Выход**: матрица размера N×40

Таблица 1. Операции по предобработке исходных аудиовизуальных данных

Отметим, что сформированные таким образом визуальные признаки не тождественны контурам рта (которые и подлежат реконструкции). Для того чтобы восстановить из визуальных признаков контур рта, следует выполнить операции, обратные Z-нормализации и нормировке по размерам рта.

Ниже приведены дополнительные пояснения для некоторых операций.

Удаление тишины в начале и в конце записи. Использован метод на основе энергии сигнала: отбрасываются те участки в начале и в конце фразы, где энергия меньше 5% от средней по всей фразе.

Выделение звуковых признаков. В качестве базовых признаков используются 13 мел-частотных кепстральных коэффициентов (Mel-Frequency Cepstral Coefficients, MFCC) и логарифм энергии (logE). На 10 мс сигнала приходится один вектор признаков.

Оценка производных. Для описания динамики сигнала по временной шкале вычисляются первая разность (Δ, соответствует первой производной) и вторая разность (ΔΔ, соответствует второй производной). Эти признаки добавляются к исходному вектору. В результате получается вектор, содержащий $3 \cdot (13 + 1) = 42$ элемента.

Z-нормализация. Этот прием предполагает центрирование относительно среднего и нормирование каждого элемента векторов признаков на среднеквадратическое отклонение (СКО). В контексте обработки MFCC Z-нормализация носит название нормализации среднего значения и дисперсии кепстра (Cepstral Mean and Variance Normalization, CMVN) [23].

Детектирование лица. Поиск лица на изображении выполняется с помощью детектора лиц, входящего в состав библиотеки компьютерного зрения OpenCV [24].

Детектирование опорных точек и выделение точек рта. Поиск 68 опорных точек лица осуществляется при помощи алгоритма на основе каскада регрессоров [25], реализованного в библиотеке Dlib [26]. Из этих 68 опорных точек 20 описывают контур рта, которые и выбираются для дальнейшей обработки.

Нормировка опорных точек. Компенсация поворота и масштаба лица осуществляется на основе положений центров глаз. Кроме того, для каждой фразы выполняется нормировка размера рта. Ширина рта вычисляется как медианное значение (0,5-квантиль) расстояния между самой левой и самой правой опорными точками рта по заданной фразе. В качестве высоты рта принимается 0,05-квантиль расстояния между верхней и нижней опорными точками. Опорные точки рта нормируются на полученные таким образом значения ширины и высоты.

Для поддержания синхронности данных на этапе предобработки каждому отсчету и каждому исходному кадру присваиваются временны́е метки (timestamps), которые рассчитываются исходя из известной частоты дискретизации. На каждом этапе предобработки обеспечивается согласованность временных меток.

Выравнивание скоростей потоков данных осуществляется с помощью интерполяции на основе кубических сплайнов. Именно на этом этапе временные метки для звуковой и визуальной составляющей сводятся в единую шкалу, относительно которой и вычисляются значения интерполирующих функций.

Для компенсации постоянного сдвига, изначально присутствующего в исходных данных, достаточно просто добавить его к соответствующим временным меткам звуковой составляющей сигнала перед интерполяцией. В данной работе использовано значение, равное 15 мс (звуковая составляющая опережает визуальную).

Построение моделей реконструкции

Исходные данные представлены в виде матриц $\mathbf{X}$ и $\mathbf{Y}$, имеющих размеры $N \times D_{\mathbf{X}}$ и $N \times D_{\mathbf{Y}}$ соответственно:

$$\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x}_N)^T, \qquad \mathbf{Y} = (\mathbf{y_1}, \mathbf{y_2}, \dots, \mathbf{y}_N)^T.$$

где $\mathbf{x}_i$ и $\mathbf{y}_i$ – векторы звуковых и визуальных признаков.

Задача реконструкции сводится к построению функции $\mathcal{F}(\mathbf{x}) \equiv \hat{\mathbf{y}} \approx \mathbf{y}$, аппроксимирующей $\mathbf{y}$ по $\mathbf{x}$. Предложенный в данной работе подход к решению этой задачи включает три этапа:

1. кластеризация звуковых признаков (необязательная);
2. проекция исходных данных на собственные подпространства;
3. построение регрессионной модели для полученных проекций.

Если была выполнена кластеризация звуковых признаков, то две последующие операции выполняются отдельно для каждого кластера.

Проекция исходных данных на собственные подпространства. Эта операция реализована в двух вариантах: на основе PCA и PLS. В обоих случаях исходные данные представляются в следующей форме:

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

$$\mathbf{Y} = \mathbf{T_Y}\mathbf{P_Y}^T + \mathbf{E_Y} = \mathbf{t}_{\mathbf{Y},1}\mathbf{p}_{\mathbf{Y},1}^T + \mathbf{t}_{\mathbf{Y},2}\mathbf{p}_{\mathbf{Y},2}^T + \cdots + \mathbf{t}_{\mathbf{Y},q}\mathbf{p}_{\mathbf{Y},q}^T + \mathbf{E_Y}.$$

Здесь использованы следующие обозначения:

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

$\mathbf{T_Y} = (\mathbf{t}_{\mathbf{Y},1}\ \mathbf{t}_{\mathbf{Y},2}\ \dots\ \mathbf{t}_{\mathbf{Y},q})$ – матрица счетов для $\mathbf{Y}$ $(N \times q)$;

$\mathbf{P_X} = (\mathbf{p}_{\mathbf{X},1}\ \mathbf{p}_{\mathbf{X},2}\ \ldots\ \mathbf{p}_{\mathbf{X},p})$ – матрица нагрузок для $\mathbf{X}$ ($D_\mathbf{X} \times p$);
$\mathbf{P_Y} = (\mathbf{p}_{\mathbf{Y},1}\ \mathbf{p}_{\mathbf{Y},2}\ \ldots\ \mathbf{p}_{\mathbf{Y},q})$ – матрица нагрузок для $\mathbf{Y}$ ($D_\mathbf{Y} \times q$);
$\mathbf{E_X}$ – матрица остатков для $\mathbf{X}$ ($N \times D_\mathbf{X}$);
$\mathbf{E_Y}$ – матрица остатков для $\mathbf{Y}$ ($N \times D_\mathbf{Y}$);
p, q – порядки моделей (в случае PLS $p = q$).

Векторы счетов $\mathbf{t}_{\mathbf{X},j}$ и $\mathbf{t}_{\mathbf{Y},k}$ представляют собой проекции исходных данных на направления $\mathbf{w}_{\mathbf{X},j} \in \mathbb{R}^{D\mathbf{x}}$ и $\mathbf{w}_{\mathbf{Y},k} \in \mathbb{R}^{D\mathbf{Y}}$. Эти векторы можно объединить в матрицы весов $\mathbf{W_X} = (\mathbf{w}_{\mathbf{X},1}\ \mathbf{w}_{\mathbf{X},2}\ \ldots\ \mathbf{w}_{\mathbf{X},p})$ и $\mathbf{W_Y} = (\mathbf{w}_{\mathbf{Y},1}\ \mathbf{w}_{\mathbf{Y},2}\ \ldots\ \mathbf{w}_{\mathbf{Y},q})$. В случае PCA $\mathbf{w}_{\mathbf{X},j}$ и $\mathbf{w}_{\mathbf{Y},k}$ совпадают с $\mathbf{p}_{\mathbf{X},j}$ и $\mathbf{p}_{\mathbf{Y},k}$ с точностью до масштаба (равного квадратному корню из соответствующих собственных чисел).

Несмотря на то, что PCA и PLS предполагают построение модели в одной и той же форме, их различия весьма существенны.

В случае PCA решается задача максимизации выборочных дисперсий в рамках каждого из наборов исходных данных:

$$\mathbf{w_X} = \operatorname{argmax}_{\mathbf{w_X}} \mathbf{w_X}^T \mathbf{X}^T \mathbf{X} \mathbf{w_X}, \qquad \text{где } \|\mathbf{w_X}\|^2 = 1,$$
$$\mathbf{w_Y} = \operatorname{argmax}_{\mathbf{w_Y}} \mathbf{w_Y}^T \mathbf{Y}^T \mathbf{Y} \mathbf{w_Y}, \qquad \text{где } \|\mathbf{w_Y}\|^2 = 1.$$

Векторы $\mathbf{w_X}$ и $\mathbf{w_Y}$ получаются как решения двух независимых задач на собственные значения:

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

В случае PLS максимизируется выборочная ковариация:

$$\{\mathbf{w_X}, \mathbf{w_Y}\} = \operatorname{argmax}_{\mathbf{w_X},\mathbf{w_Y}} \mathbf{w_X}^T \mathbf{X}^T \mathbf{Y} \mathbf{w_Y} = \operatorname{argmax}_{\mathbf{w_X},\mathbf{w_Y}} \mathbf{t_X}^T \mathbf{t_Y}, \qquad \text{где } \|\mathbf{w_X}\|^2 = \|\mathbf{w_Y}\|^2 = 1.$$

Задача на собственные значения имеет следующий вид (отметим, что в этом случае набор собственных чисел λ является общим для $\mathbf{X}$ и $\mathbf{Y}$):

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

Эта задача эффективно решается с помощью алгоритма NIPALS (NonLinear Iterative Partial Least Squares) [21, 22], который использован в данной работе. Более подробное описание метода регрессии PLS и алгоритма NIPALS в тех же обозначениях, которые использованы здесь, приведено в [27].

Существенным отличием PCA от PLS является то, что в первом случае выполняется независимый анализ для исходных наборов данных $\mathbf{X}$ и $\mathbf{Y}$, в то время как PLS учитывает ковариационные связи между ними. Этот факт говорит в пользу PLS.

Проекция исходных данных на собственные подпространства позволяет снизить размерность пространства признаков, обеспечить устойчивость решения регрессионной задачи (например, решить проблему коллинеарности [22]) и использовать для построения регрессионной модели только «существенные» (в смысле дисперсии или ковариации) составляющие исходных данных.

Построение регрессионной модели. Используемая в данной работе модель является линейной, т.е. $\mathcal{F}(\mathbf{x}) = \mathbf{R}^T\mathbf{x}$. Матрица регрессии $\mathbf{R}$ подбирается таким образом, чтобы реконструированные визуальные признаки $\widehat{\mathbf{Y}}$ аппроксимировали истинные значения $\mathbf{Y}$:

$$\mathbf{XR} = \widehat{\mathbf{Y}} \approx \mathbf{Y}.$$

В случае PCA используется метод множественной линейной регрессии (Multiple Linear Regression, MLR):

$$\mathbf{R} = \mathbf{W_X}(\mathbf{T_X}^+\mathbf{T_Y})\mathbf{W_Y}^T,$$

где $\mathbf{A}^+ \stackrel{\text{def}}{=} (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ – псевдообратная матрица.

Здесь левое и правое умножение на матрицы весов $\mathbf{W_X}$ и $\mathbf{W_Y}^T$ фактически реализует прямое и обратное преобразование Карунена–Лоэва (Karhunen–Loève transform, KLT) в пространстве звуковых и визуальных признаков соответственно.

В случае регрессии PLS матрица $\mathbf{R}$ вычисляется следующим образом [21]:

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

Здесь важно отметить следующее. Так как матрица $\mathbf{R}$ имеет размерность $D_\mathbf{X} \times D_\mathbf{Y}$ (в нашем случае 42× 40), вычисление функции $\mathcal{F}(\mathbf{x}) = \mathbf{R}^T\mathbf{x}$ требует $D_\mathbf{X} \cdot D_\mathbf{Y} = 42 \cdot 40 = 1680$ операций умножения. С другой стороны, матрицу $\mathbf{R}$ можно представить как произведение трех матриц, имеющих размерности $D_\mathbf{X} \times p$, $p \times q$ и $q \times D_\mathbf{Y}$. Если число компонент равно, например, пяти, т.е. $p = q = 5$ (ниже показано, что в случае регрессии PLS такое значение достижимо), то вычисление $\mathcal{F}(\mathbf{x})$ реализуется за

$$D_\mathbf{X} \cdot p + p \cdot q + q \cdot D_\mathbf{Y} = 42 \cdot 5 + 5 \cdot 5 + 5 \cdot 40 = 435$$

операций умножения. Это почти в четыре раза меньше, чем при непосредственном умножении на матрицу $\mathbf{R}$. Таким образом, при правильной реализации меньшее количество компонент позволяет значительно снизить вычислительные затраты.

Кластеризация звуковых признаков. Следует отметить, что в некоторых случаях точности линейной регрессионной модели может быть недостаточно для моделирования скрытых связей между модаль-

ностями. С другой стороны, нелинейные модели сложны, плохо поддаются интерпретации и требуют значительных вычислительных затрат. В этой связи для повышения точности регрессионной модели предлагается промежуточное решение, которое можно назвать «кусочно-линейной» моделью. Для этого звуковые признаки разбиваются на кластеры при помощи алгоритма K-средних. Векторы визуальных признаков группируются во множества, соответствующие построенным «звуковым» кластерам. Затем для каждого кластера формируется отдельная регрессионная модель.

Такой подход обоснован тем, что речь состоит из структурных элементов – фонем и визем. С другой стороны, соответствие построенных кластеров фонемам и виземам никоим образом не предполагается.

Экспериментальные исследования

Для проведения экспериментальных исследований представленные выше алгоритмы предобработки, регрессии PLS, PCA, MLR и кластеризации K-средних реализованы в среде математического пакета MATLAB. Для удаления шума из исходных аудиозаписей использовано приложение SoX [28]. Детектирование лица и опорных точек выполнено с помощью функций, реализованных в библиотеках OpenCV [24] и Dlib [26]. Также использована процедура выделения MFCC [29].

В рамках экспериментальных исследований выполнена реконструкция контура рта по голосу. Прежде всего, это связано с тем, что звуковая составляющая содержит больше информации о структуре речевого сигнала. Этот факт обусловлен тем, что значительная часть речевого тракта человека не отображена на видеозаписи. Кроме того, визуальные признаки, выраженные в форме набора опорных точек, хорошо поддаются сравнению и интерпретации.

Для проведения экспериментов реализованы четыре варианта представленного выше подхода:

1. PCA и MLR (далее «PCA»);
2. регрессия PLS (далее «PLS»);
3. PCA и MLR с предварительной кластеризацией (далее «PCA + кластеризация»);
4. регрессия PLS с предварительной кластеризацией (далее «PLS + кластеризация»).

Исследования выполнены на аудиовизуальной англоязычной базе VidTIMIT [30, 31]. Она была разделена на подмножества «train» (обучающее), «validation» (проверочное) и «test» (тестовое). В табл. 2 приведены характеристики этих подмножеств и базы в целом. В длительность речи не включены участки тишины, удаленные на этапе предобработки. Отметим, что в разных множествах содержатся записи разных людей (дикторов).

	Train	Validation	Test	VidTIMIT
Количество дикторов	14 мужчин 11 женщин	5 мужчин 4 женщины	5 мужчин 4 женщины	24 мужчины 19 женщин
Количество фраз	250	90	90	430
Длительность речи, мин	12	5	5	22

Таблица 2. Характеристики базы VidTIMIT и ее подмножеств

Построение регрессионных моделей. Для построения регрессионных моделей используются признаки, полученные с помощью процедуры, описанной в табл. 1. Для отслеживания процесса обучения использована нормированная среднеквадратическая ошибка (Normalized Root-Mean-Square Error, NRMSE):

$$NRMSE = \sqrt{\frac{\sum_{k=1}^{K}\|\mathbf{y}_k-\hat{\mathbf{y}}_k\|^2}{\sum_{k=1}^{K}\|\mathbf{y}_k-\bar{\mathbf{y}}\|^2}}. \quad (1)$$

Напомним, что NRMSE вычисляется для визуальных признаков, которые, как было отмечено выше, не следует отождествлять с контуром рта. Кроме того, единого подхода к нормировке среднеквадратической ошибки не существует. В некоторых случаях выполняют нормировку на размах выборки или на среднее значение (в (1) сделано именно это), а иногда и вовсе отказываются от нее. Таким образом, NRMSE не дает полного представления о качестве реконструкции. С другой стороны, эта метрика позволяет контролировать процесс обучения и выявлять случаи переобучения (overfitting) путем сравнения значений NRMSE на множествах «train» и «validation». Ниже для количественной оценки качества реконструкции представлена другая метрика (2), которая вычисляется уже относительно контура рта.

Процесс построения регрессионных моделей иллюстрируют следующие данные:

– графики собственных чисел для PCA и PLS. Чем быстрее убывают собственные числа, тем меньше компонент достаточно для последующего анализа;
– кривые обучения (Learning Curves), отражающие зависимость NRMSE от объема обучающей выборки на множествах «train» и «validation». Кривые обучения позволяют выявлять случаи переобучения и определять, какой минимальный объем данных достаточен для построения хорошей модели;
– гистограммы распределения исходных данных по кластерам.

На рис. 1 представлены графики собственных чисел и кривые обучения, соответствующие варианту «PCA». На рис. 1, а, вертикальной чертой обозначено количество главных компонент, использованное для построения регрессионной модели.

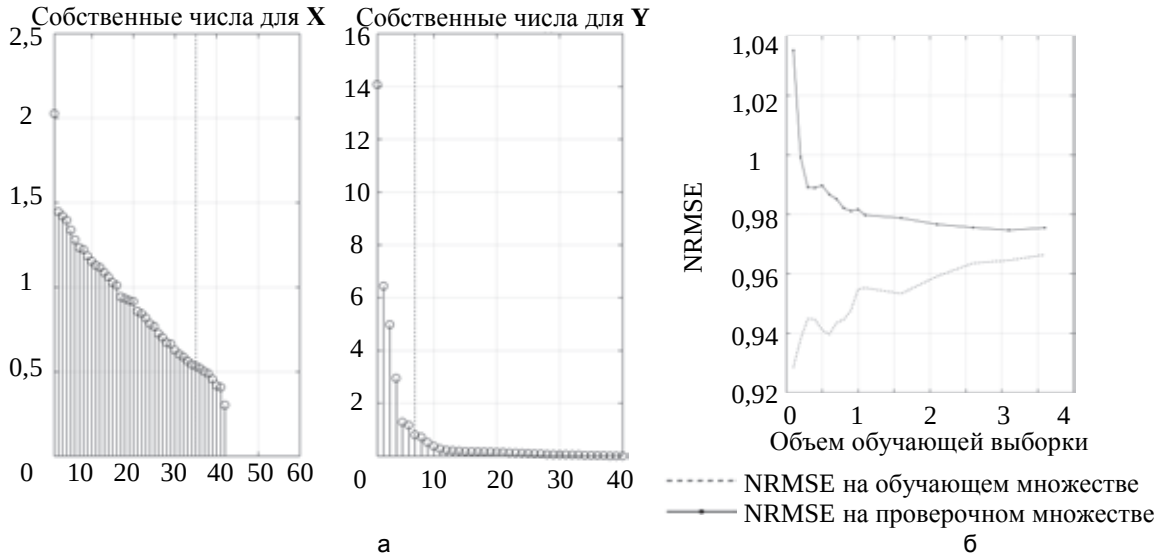


Рис. 1. Собственные числа (а) и кривые обучения (б), полученные с помощью варианта «PCA»

На рис. 2 показаны графики собственных чисел и кривые обучения, соответствующие варианту «PLS». Обратим внимание, что в этом случае имеется только один общий набор собственных чисел. Для анализа выбрано пять компонент. Заметим, что алгоритм NIPALS предполагает последовательное вычисление компонент в порядке убывания соответствующих им собственных чисел.

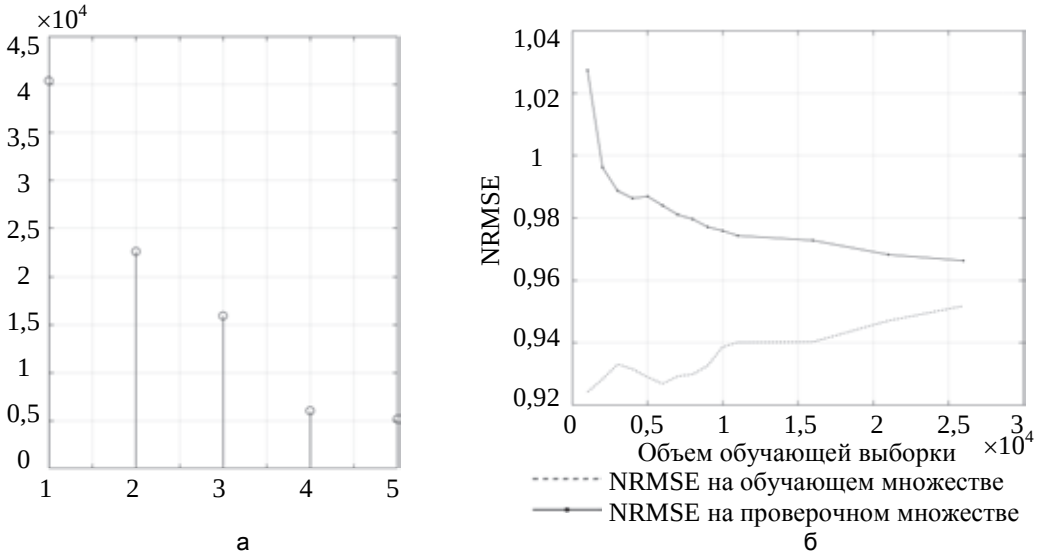


Рис. 2. Собственные числа (а) и кривые обучения (б), соответствующие варианту «PLS»

На рис. 3 представлены графики собственных чисел и кривые обучения, соответствующие варианту «PCA + кластеризация». Кроме того, показано распределение векторов данных по кластерам. Это важно для того, чтобы убедиться в отсутствии «маленьких» кластеров, что может привести к плохим регрессионным моделям. Для каждого кластера также приведены соответствующие собственные числа. Показаны только графики для звуковой составляющей (в случае визуальной составляющей они не представляют особого интереса).

На рис. 4 представлена гистограмма распределения данных по кластерам, собственные числа и кривые обучения для моделей, полученных с помощью варианта «PLS + кластеризация». Рис. 4 аналогичен рис. 3, за исключением того, что собственные числа на графиках являются общими для звуковых и визуальных признаков.

Полученные данные позволяют сделать следующие выводы.

– Объем обучающей выборки (всего 12 мин речи) достаточен для построения регрессионных моделей любым из предложенных выше способов. Такой вывод можно сделать на основе анализа кривых обучения: при достижении максимального значения объема обучающей выборки NRMSE на множестве «train» приближается к NRMSE на множестве «validation». Это говорит о том, что модели не переобучены (даже в тех случаях, когда используется кластеризация).
– Визуальные признаки могут быть достаточно точно смоделированы при помощи всего семи компонент, что видно из графика собственных чисел для визуальных признаков (рис. 1, а).
– Для моделирования звуковых признаков необходимо большое количество главных компонент (рис. 1, а). Это объясняется тем, что MFCC по своей природе уже декоррелированы и несут информацию о речевом сигнале в достаточно сжатой форме.

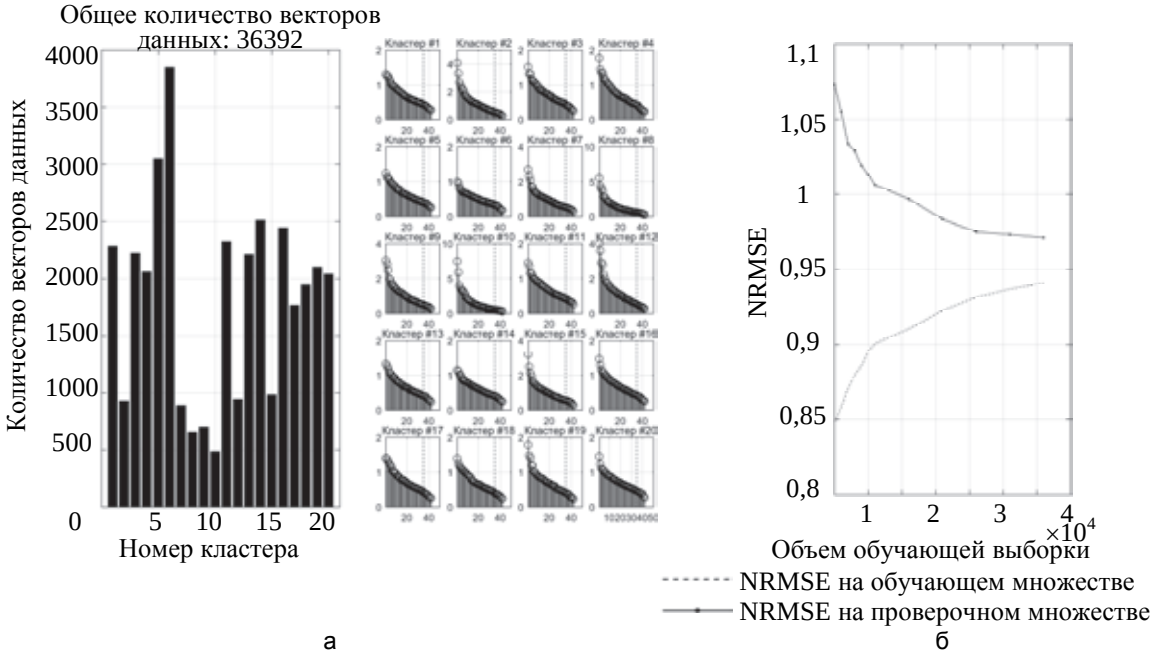


Рис. 3. Распределение данных по кластерам, собственные числа для каждого кластера звуковой составляющей (а) и кривые обучения (б), полученные с помощью варианта «PCA + кластеризация»

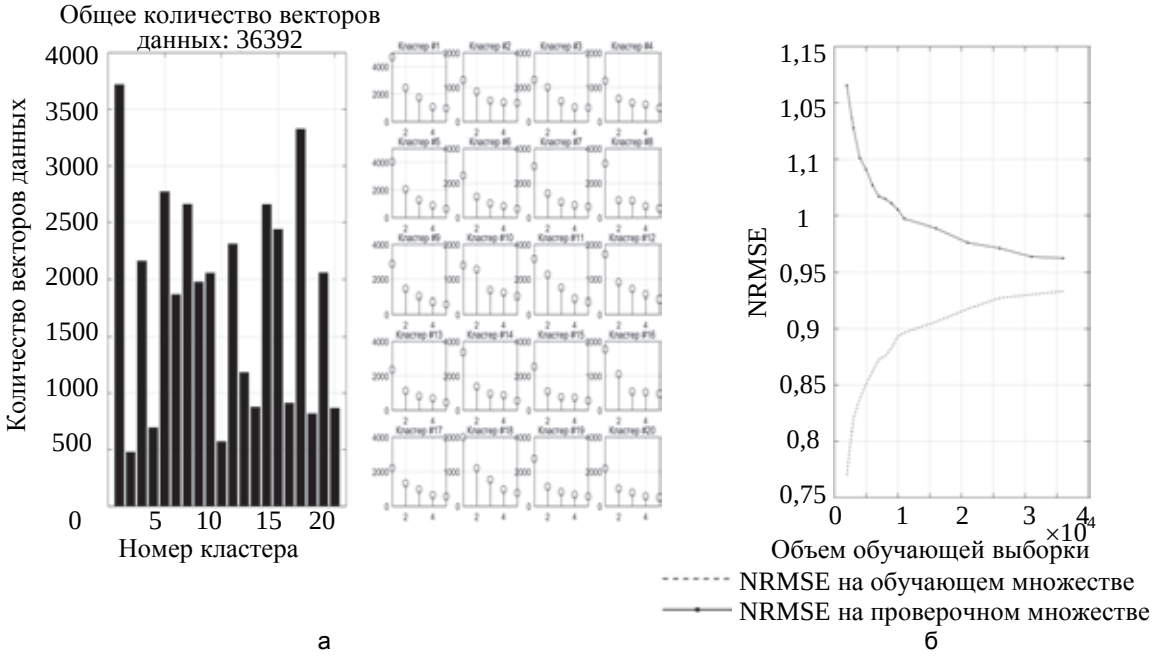


Рис. 4. Распределение данных по кластерам, собственные числа для каждого кластера (а) и кривые обучения (б), полученные с помощью варианта «PLS + кластеризация»

Оценка качества реконструкции контура рта. Представленные выше результаты позволяют сделать вывод о том, что полученные модели не переобучены, а также оценить количество главных компо-

нент, необходимых для построения регрессионной модели. Однако для оценки качества реконструкции контура рта необходима другая метрика, отличная от NRMSE.

Прежде всего, каждый контур рта восстанавливается из соответствующего вектора визуальных признаков при помощи операций, обратных Z-нормализации, и нормировке по размерам рта. Необходимые для этого средние значения, СКО и средние размеры рта (ширина W и высота H) вычислены по обучающему множеству («train»). Как было сказано выше, исходный контур рта описывается двадцатью опорными точками на плоскости $\left(c_{x_i}^{(k)}, c_{y_i}^{(k)}\right)$, где $k = 1, \dots, K$ – номер образца в наборе данных, $i = 1, \dots, 20$ – номер опорной точки. Соответствующие восстановленные по голосу точки обозначим как $\left(\hat{c}_{x_i}^{(k)}, \hat{c}_{y_i}^{(k)}\right)$.

Определим ошибку реконструкции для k-го контура рта как среднее расстояние между соответствующими исходными и реконструированными точками, нормированное на полученное из обучающей выборки среднее значение ширины рта W:

$$\varepsilon_k = \frac{1}{W \cdot 20} \sum_{i=1}^{20} \sqrt{\left(c_{x_i}^{(k)} - \hat{c}_{x_i}^{(k)}\right)^2 + \left(c_{y_i}^{(k)} - \hat{c}_{y_i}^{(k)}\right)^2}. \quad (2)$$

Нормировка на W в (2) нужна для того, чтобы привести значения ошибки реконструкции к некоторой интерпретируемой шкале. В данном случае можно говорить о том, что ε_k измеряется в долях от средней ширины рта.

Для значений ε_k, полученных на множестве «test», вычислены среднее $\bar{\varepsilon}$, медиана (0,5-квантиль) $\varepsilon_{0,5}$ и 0,95-квантиль $\varepsilon_{0,95}$. Последнее значение позволяет оценить ошибку реконструкции в «плохих» случаях, не принимая во внимание присутствующие в тестовой выборке выбросы. Полученные данные представлены в табл. 3. Для вариантов «PCA» и «PCA + кластеризация» отдельно для звуковой и визуальной составляющих указано количество компонент, использованных для построения регрессионной модели. Наилучшие результаты показал вариант «PLS + кластеризация».

На рис. 5, а, для варианта «PCA + кластеризация» показана визуализация значений $\bar{\varepsilon}$, $\varepsilon_{0,5}$ и $\varepsilon_{0,95}$ относительно контура рта, что позволяет получить представление о достигнутой точности реконструкции. На рис. 5, б, показана гистограмма распределения ошибки ε_k для того же варианта.

На рис. 5 представлены примеры реконструированного контура рта для варианта «PLS + кластеризация». Данные также взяты из множества «test».

Метод реконструкции	Ошибка реконструкции		
	среднее $\bar{\varepsilon}$	медиана $\varepsilon_{0,5}$	0,95-квантиль $\varepsilon_{0,95}$
PCA (5/7 компонент)	0,0252	0,0230	0,0470
PCA + кластеризация (5/7 компонент)	0,0248	0,0225	0,0466
PCA (35/7 компонент)	0,0244	0,0224	0,0457
PCA + кластеризация (35/7 компонент)	0,0241	0,0220	0,0444
PLS (5 компонент)	0,0240	0,0220	0,0447
PLS + кластеризация (5 компонент)	0,0238	0,0218	0,0439

Таблица 3. Среднее, медиана и 0,95-квантиль ошибки реконструкции ε_k, вычисленные на множестве «test» для регрессионных моделей, полученных различными способами

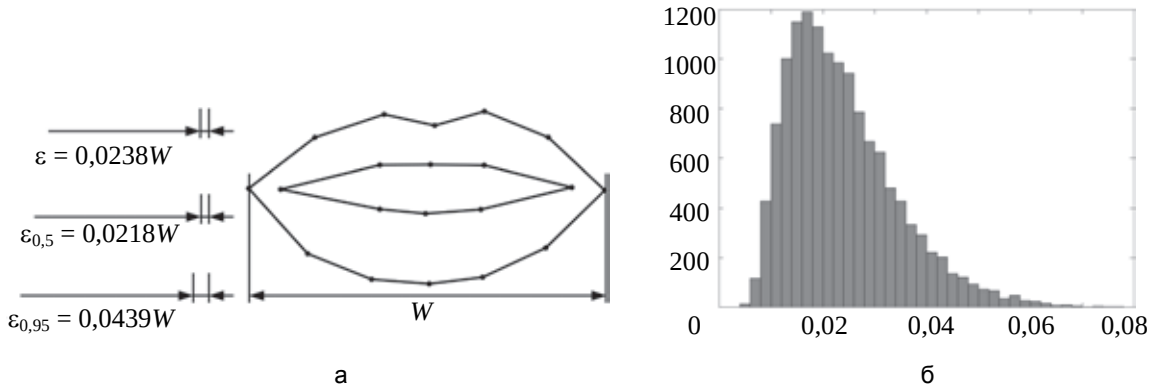
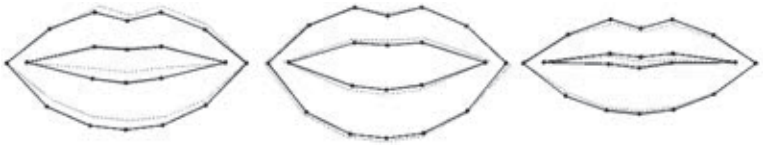


Рис. 5. Визуализация ошибки реконструкции (а) и гистограмма распределения значений ε_k (б) для варианта «PCA + кластеризация»



а б в

Рис. 6. Примеры реконструкции контура рта на тестовой выборке с помощью варианта «PLS + кластеризация». Сплошной линией обозначен реконструированный по голосу контур, а пунктирной – «настоящий» контур рта, снятый с видеозаписи. Показаны результаты для полуоткрытого (а), открытого (б) и закрытого (в) рта

На основе полученных результатов можно сделать следующие выводы.

– Как количественная (табл. 3 и рис. 5), так и качественная (рис. 6) оценки результатов реконструкции контура рта позволяет сделать вывод о том, что предложенный подход успешно решает поставленную задачу реконструкции контура рта по голосу.
– Все рассмотренные варианты предложенного подхода дают близкие по качеству результаты. Однако PLS позволяет использовать гораздо меньше компонент при несколько лучшем (по сравнению с PCA и MLR) качестве реконструкции. Как показано выше, это позволяет существенно сократить вычислительные затраты. Такой результат является следствием того, что PLS, в отличие от PCA, моделирует взаимосвязи между исходными наборами данных.
– Кластеризация звуковых признаков позволяет несколько снизить ошибку реконструкции.
– Значения $\varepsilon_{0,95}$ и гистограмма распределения значений ε_k (рис. 5, б) позволяют говорить о том, что даже в «плохих» случаях ошибка реконструкции остается в разумных пределах.

Заключение

В работе рассмотрена задача выявления скрытых связей между звуковой и визуальной модальностями речевого сигнала и их взаимной реконструкции. Предложенный подход основан на проекционных и регрессионных методах – PCA, MLR и регрессии PLS. Кроме того, для повышения точности реконструкции использована кластеризация на основе алгоритма K-средних.

Эксперименты по реконструкции контура рта по голосу проведены на аудиовизуальной англоязычной базе VidTIMIT. Представлены варианты реализации предложенного подхода на основе PCA и регрессии PLS с предварительной кластеризацией звуковых признаков и без нее (всего четыре варианта). Как количественная (объективная), так и качественная (субъективная) оценки подтвердили работоспособность предложенного подхода; наилучшие результаты показала реализация на основе регрессии PLS с кластеризацией.

Несмотря на то, что кластеризация несколько улучшает качество реконструкции, она несколько усложняет систему и увеличивает вычислительные затраты. По этой причине во многих случаях обоснован отказ от этого этапа в пользу простой линейной модели.

Одним из важных преимуществ предложенного подхода является отсутствие потребности в объемной обучающей выборке (достаточно всего 12 минут речи), что также подтверждено результатами экспериментов. Следует отметить, что по сравнению с методами PCA и MLR регрессия PLS позволяет добиться более высокого качества реконструкции при меньшем количестве компонент. При этом показано, что использование меньшего количества компонент позволяет существенно снизить вычислительные затраты.

На основе предложенного подхода могут быть разработаны разнообразные решения: бимодальные биометрические системы, управляемые голосом виртуальные двойники («аватары»), системы контроля доступа к мобильным устройствам и другие решения в области аудиовизуальных человеко-машинных интерфейсов.

Так как исследования проводились только на одной базе, неизвестно, насколько хорошо построенные модели будут работать в произвольных условиях. Однако поскольку обучение системы возможно без объемной обучающей выборки, проверка пригодности предложенных решений для той или иной практической задачи не требует больших трудозатрат.

Одним из возможных путей повышения качества реконструкции и надежности предложенного решения является замена MFCC на другой экстрактор голосовых признаков, хорошо работающий в сложных условиях. Такой экстрактор может обучаться совместно с регрессионной моделью подобно тому, как это делается в методе «глубокого канонического корреляционного анализа» (Deep CCA) [32], однако следует понимать, что в этом случае может потребоваться обучающая выборка гораздо большего объема. Кроме того, предложенная модель не учитывает тот факт, что некоторые звуки, играющие важную роль в восприятии смысла предложения (например, взрывные согласные), занимают лишь малую долю всей

временной шкалы. Отразить эту особенность в модели можно, если делать выборку звуковых признаков с разной частотой, определяемой характером самого речевого сигнала.

Литература

1. Иванько Д.В., Карпов А.А. Анализ перспектив применения высокоскоростных камер для распознавания динамической видеоинформации // Труды СПИИРАН. 2016. № 1. С. 98–113. doi: 10.15622/SP.44.7
2. McGurk H., MacDonald J. Hearing lips and seeing voices // Nature. 1976. V. 264. N 5588. P. 746–748.
3. Atrey P.K., Hossain M.A., El Saddik A., Kankanhalli M.S. Multimodal fusion for multimedia analysis: a survey // Multimedia Systems. 2010. V. 16. N 6. P. 345–379. doi: 10.1007/s00530-010-0182-0
4. Nefian A.V., Liang L., Pi X. et al. A coupled HMM for audio-visual speech recognition // Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing. 2002. V. 2. P. 2013–2016. doi: 10.1109/ICASSP.2002.5745027
5. Карпов А.А. Реализация автоматической системы многомодального распознавания речи по аудио- и видеоинформации // Автоматика и телемеханика. 2014. № 12. С. 125–138.
6. Pachoud S., Gong S., Cavallaro A. Space-time audio-visual speech recognition with multiple multi-class probabilistic support vector machines // Proc. Auditory-Visual Speech Processing AVSP. Norwich, UK, 2009. P. 155–160.
7. Hammami I., Mercies G., Hamouda A. The Kohonen map for credal fusion of heterogeneous data // Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Milan, Italy, 2015. P. 2947–2950. doi: 10.1109/IGARSS.2015.7326433
8. Hochreiter S., Schmidhuber J. Long short-term memory // Neural Computation. 1997. V. 9. N 8. P. 1735–1780. doi: 10.1162/neco.1997.9.8.1735
9. Jaeger H. The «echo state» approach to analysing and training recurrent neural networks - with an erratum note // GMD Technical Report 148, German National Research Center for Information Technology, 2001. 13 p.
10. LeCun Y. et al. Gradient-based learning applied to document recognition // Proceedings of the IEEE. 1998. V. 86. N 11. P. 2278–2324. doi: 10.1109/5.726791
11. Hou J.-C., Wang S.S., Lai Y.H., Tsao Y., Chang H.W., Wan H.M. Audio-visual speech enhancement based on multimodal deep convolutional neural network // ArXiv Prepr. ArXiv170310893. 2017.
12. Noda K., Yamaguchi Y., Nakadai K., Okuno H.G., Ogata T. Audio-visual speech recognition using deep learning // Applied Intelligence. 2015. V. 42. N 4. P. 722–737. doi: 10.1007/s10489-014-0629-7
13. Ren J., Hu Y., Tai Y.W. et al. Look, listen and learn - a multimodal LSTM for speaker identification // Proc. 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016. P. 3581–3587.
14. 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.
15. 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
16. Liu M., Wang R., Huang Z., Shan S., Chen X. Partial least squares regression on grassmannian manifold for emotion recognition // Proc. 15th ACM on Int. Conf. on Multimodal Interaction. Sydney, Australia, 2013. P. 525–530. doi: 10.1145/2522848.2531738
17. 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
18. Sargin M.E., Yemez Y., Erzin E., Tekalp A.M. Audiovisual synchronization and fusion using canonical correlation analysis

References

1. Ivanko D.V., Karpov A.A. An analysis of perspectives for using high-speed cameras in processing dynamic video information. *SPIIRAS Proceedings*, 2016, no. 1, pp. 98–113. doi: 10.15622/SP.44.7 (In Russian)
2. McGurk H., MacDonald J. Hearing lips and seeing voices. *Nature*, 1976, vol. 264, no. 5588, pp. 746–748.
3. Atrey P.K., Hossain M.A., El Saddik A., Kankanhalli M.S. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 2010, vol. 16, no. 6, pp. 345–379. doi: 10.1007/s00530-010-0182-0
4. Nefian A.V., Liang L., Pi X. et al. A coupled HMM for audio-visual speech recognition. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2002, vol. 2, pp. 2013–2016. doi: 10.1109/ICASSP.2002.5745027
5. Karpov A. An automatic multimodal speech recognition system with audio and video information. *Automation and Remote Control*, 2014, vol. 75, no. 12, pp. 2190–2200. doi: 10.1134/S000511791412008X
6. Pachoud S., Gong S., Cavallaro A. Space-time audio-visual speech recognition with multiple multi-class probabilistic support vector machines. *Proc. Auditory-Visual Speech Processing AVSP*. Norwich, UK, 2009, pp. 155–160.
7. Hammami I., Mercies G., Hamouda A. The Kohonen map for credal fusion of heterogeneous data. *Proc. IEEE International Geoscience and Remote Sensing Symposium, IGARSS*. Milan, Italy, 2015, pp. 2947–2950. doi: 10.1109/IGARSS.2015.7326433
8. Hochreiter S., Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, vol. 9, no. 8, pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735
9. Jaeger H. The «echo state» approach to analysing and training recurrent neural networks - with an erratum note. *GMD Technical Report 148, German National Research Center for Information Technology*, 2001, 13 p.
10. LeCun Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, vol. 86, no. 11, pp. 2278–2324. doi: 10.1109/5.726791
11. Hou J.-C., Wang S.S., Lai Y.H., Tsao Y., Chang H.W., Wan H.M. Audio-visual speech enhancement based on multimodal deep convolutional neural network. *ArXiv Prepr*, ArXiv170310893, 2017.
12. Noda K., Yamaguchi Y., Nakadai K., Okuno H.G., Ogata T. Audio-visual speech recognition using deep learning. *Applied Intelligence*, 2015, vol. 42, no. 4, pp. 722–737. doi: 10.1007/s10489-014-0629-7
13. Ren J., Hu Y., Tai Y.W. et al. Look, listen and learn - a multimodal LSTM for speaker identification. *Proc. 30th AAAI Conference on Artificial Intelligence*. Phoenix, USA, 2016, pp. 3581–3587.
14. Kukharev G.A., Kamenskaya E.I., Matveev Y.N., Shchegoleva N.L. *Methods for Face Image Processing and Recognition in Biometric Applications*. Ed. M.V. Khitrov. St. Petersburg, Politekhnika Publ., 2013, 388 p. (In Russian)
15. 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, pp. 21–29. doi: 10.1145/2512530.2512532
16. Liu M., Wang R., Huang Z., Shan S., Chen X. Partial least squares regression on grassmannian manifold for emotion recognition. *Proc. 15th ACM on Int. Conf. on Multimodal Interaction*. Sydney, Australia, 2013, pp. 525–530. doi: 10.1145/2522848.2531738
17. 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, pp. 684–691. doi: 10.1109/CVPR.2013.94
18. Sargin M.E., Yemez Y., Erzin E., Tekalp A.M. Audiovisual

// IEEE Transactions on Multimedia. 2007. V. 9. N 7. P. 1396–1403. doi: 10.1109/TMM.2007.906583
19. Sigg C., Fischer B., Ommer B., Roth V., Buhmann J. Nonnegative CCA for audiovisual source separation // Proc. 17th IEEE Int. Workshop on Machine Learning for Signal Processing. Thessaloniki, Greece, 2007. P. 253–258. doi: 10.1109/MLSP.2007.4414315
20. Lee J.-S., Ebrahimi T. Two-level bimodal association for audio-visual speech recognition // Lecture Notes in Computer Science. 2009. V. 5807. P. 133–144. doi: 10.1007/978-3-642-04697-1_13
21. 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
22. Эсбенсен К. Анализ многомерных данных. Черноголовка: ИПХФ РАН, 2005. 160 с.
23. Prasad N.V., Umesh S. Improved cepstral mean and variance normalization using Bayesian framework // Proc. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. 2013. P. 156–161. doi: 10.1109/ASRU.2013.6707722
24. OpenCV Library [Электронный ресурс]. URL: http://opencv.org (дата обращения: 20.01.2018).
25. Kazemi V., Sullivan J. One millisecond face alignment with an ensemble of regression trees // Proc. IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014. P. 1867–1874. doi: 10.1109/CVPR.2014.241
26. dlib C++ Library [Электронный ресурс]. URL: http://dlib.net (дата обращения: 20.01.2018).
27. 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. doi: 10.17586/2226-1494-2015-15-5-886-892
28. SoX - Sound eXchange. HomePage [Электронный ресурс]. URL: http://sox.sourceforge.net (дата обращения: 09.09.2017).
29. 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: 20.01.2018).
30. The VidTIMIT Audio-Video Database [Электронный ресурс]. URL: http://conradsanderson.id.au/vidtimit/ (дата обращения: 20.01.2018).
31. Sanderson C., Lovell B.C. Multi-region probabilistic histograms for robust and scalable identity inference // Lecture Notes in Computer Science. 2009. V. 5558. P. 199–208. doi: 10.1007/978-3-642-01793-3_21
32. Benton A., Khayrallah H., Gujral B., Reisinger D.A., Zhang S., Arora R. Deep generalized canonical correlation analysis // arXiv:1702.02519. 2017. 14 p.

Авторы

Олейник Андрей Леонидович – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 57190279071, ORCID ID: 0000-0001-9425-2572, aoleinik@corp.ifmo.ru

synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia*, 2007, vol. 9, no. 7, pp. 1396–1403. doi: 10.1109/TMM.2007.906583
19. Sigg C., Fischer B., Ommer B., Roth V., Buhmann J. Nonnegative CCA for audiovisual source separation. *Proc. 17th IEEE Int. Workshop on Machine Learning for Signal Processing*. Thessaloniki, Greece, 2007, pp. 253–258. doi: 10.1109/MLSP.2007.4414315
20. Lee J.-S., Ebrahimi T. Two-level bimodal association for audio-visual speech recognition. *Lecture Notes in Computer Science*, 2009, vol. 5807, pp. 133–144. doi: 10.1007/978-3-642-04697-1_13
21. De Bie T., Cristianini N., Rosipal R. Eigenproblems in pattern recognition. In: *Handbook of Geometric Computing*. Ed. E.B. Corrochano. Berlin, Springer, 2005, pp. 129–167. doi: 10.1007/3-540-28247-5_5
22. Esbensen K.H. *Multivariate Date Analysis – In Practice*. 5th ed. Oslo, Norway, CAMO Process AS, 2002, 598 p.
23. Prasad N.V., Umesh S. Improved cepstral mean and variance normalization using Bayesian framework. *Proc. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 156–161. doi: 10.1109/ASRU.2013.6707722
24. *OpenCV Library*. URL: http://opencv.org (accessed: 20.01.2018).
25. Kazemi V., Sullivan J. One millisecond face alignment with an ensemble of regression trees. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, USA, 2014, pp. 1867–1874. doi: 10.1109/CVPR.2014.241
26. *dlib C++ Library*. URL: http://dlib.net (accessed: 20.01.2018).
27. 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. (In Russian) doi: 10.17586/2226-1494-2015-15-5-886-892
28. *SoX - Sound eXchange*. HomePage. URL: http://sox.sourceforge.net (accessed: 09.09.2017).
29. Wojcicki K. *Mel Frequency Cepstral Coefficient Feature Extraction*. Available at: www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab (accessed: 20.01.2018).
30. *The VidTIMIT Audio-Video Database*. URL: http://conradsanderson.id.au/vidtimit/ (accessed: 20.01.2018).
31. Sanderson C., Lovell B.C. Multi-region probabilistic histograms for robust and scalable identity inference. *Lecture Notes in Computer Science*, 2009, vol. 5558, pp. 199–208. doi: 10.1007/978-3-642-01793-3_21
32. Benton A., Khayrallah H., Gujral B., Reisinger D.A., Zhang S., Arora R. Deep generalized canonical correlation analysis. *ArXiv Prepr*, ArXiv1702.02519, 2017, 14 p.

Authors

Andrey L. Oleinik – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 57190279071, ORCID ID: 0000-0001-9425-2572, aoleinik@corp.ifmo.ru