Алгоритм визуализации речевых сигналов для интерактивного обучения правильному произношению

И.В. Губочкин

В статье предложен алгоритм визуализации речевых сигналов с использованием нелинейного отображения многомерных данных на основе метода Сэммона. Приведена модификация этого метода на случай добавления ещё одного элемента данных, а также проведён сравнительный анализ качества отображения множества реализаций фонем при выборе различных векторов признаков и метрик между ними. Показано, что главное отличие предлагаемого алгоритма состоит в возможности визуализации степени близости произношения к группе эталонных реализаций.

Abstract

The speech signal visualization algorithm using the Sammon nonlinear mapping of multidimensional data is proposed. Also it is proposed the method modification for adding yet another element of the data and a comparative analysis of the quality of displaying multiple realizations of phonemes in the selection of feature vectors and metrics between them. It is showed that the main difference of proposed algorithm is the ability to visualize the degree of pronunciation proximity to the group of reference patterns.

Введение. В настоящее время большинство систем постановки произношения основано на сравнении произнесённого слова или фонемы с некоторым эталоном. При этом произношение считается тем лучше, чем меньше расстояние между эталоном и входным сигналом.

Недостаток такого подхода заключается в том, что диктор может быть не в состоянии произнести звук, в достаточной степени похожий на эталон. Выходом из сложившейся ситуации могло бы стать сравнение с несколькими эталонами. В этом случае диктору достаточно было бы приблизить своё произношение к одному из них или к определённой доверительной области.

Данный подход позволит сократить время на обучение, поскольку диктор может приближать своё произношение к наиболее удобному для него варианту. Существующие системы постановки произношения такого типа [1, 2] требуют, в силу своих особенностей, большие объёмы обучающих данных, что не всегда достижимо на практике. Кроме того, при изучении иностранных языков обучаемый часто делает типичные ошибки в произношении определённых звуков. Механизм, который бы позволял человеку определить не только тот факт, что его произношение далеко от эталонного, но и то, какую именно ошибку он совершает, значительно повысил бы эффективность системы постановки правильного произношения.

Однако этот подход в своей реализации наталкивается на серьёзную задачу — отображение входного сигнала и эталонов таким образом, чтобы диктор мог точно определить, насколько его произношение соответствует требуемому. Ещё одна трудность состоит в том, что речевые сигналы, как правило, представляются в виде вектора признаков достаточно большой размерности (10 и более). Поэтому их непосредственное отображение на плоскости или в трёхмерном пространстве зачастую невозможно.

Выходом из данной ситуации является уменьшение размерности представляемых данных. Для выполнения этой операции в настоящее время разработан широкий класс методов. Наиболее распространёнными из них являются многомерное шкалирование [3], нелинейное отображение [4] и криволинейный компонентный анализ [5]. Основным недостатком перечисленных методов является то, что они ориентированы, прежде всего, на работу с уже сформированным множеством исходных данных и не позволяют добавлять новые данные по мере необходимости. Требуется полный пересчёт всего множества данных. Эта особенность затрудняет реализацию эффективных методов визуализации расположения входного сигнала относительно эталонов. Решению перечисленных проблем и посвящена данная работа.

Алгоритм отображения. За основу был взят широко распространённый метод Сэммона [5]. Его основным преимуществом является тот факт, что он стремится сохранить, насколько это возможно, расстояния между близлежащими точками. Это свойство особенно важно в задаче автоматической постановки правильного произношения, поскольку кластеры, соответствующие реализациям определённых фонем, должны сохранять свою компактность при отображении.

Метод Сэммона позволяет достичь минимизации критерия ошибки, который характеризует разности расстояний между точками в исходном и отображаемом пространствах. Обозначим через δ_{ij} расстояние между векторами x_i и x_j в исходном пространстве размерности n, а через d_{ij} — расстояние между векторами y_i и y_j в пространстве отображения размерности q, $q < n$. Тогда суммарная ошибка с учётом нормирующего множителя будет равна:

$$E = \frac{1}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\delta_{ij}} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{(\delta_{ij}-d_{ij})^2}{\delta_{ij}}\ , \qquad (1)$$

где N — число векторов множества входных данных X.

Для того чтобы получить оптимальное в смысле сохранения расстояния отображение, найдём минимум E по y_j для случая использования в качестве d_{ij} евклидова расстояния. Метод наискорейшего спуска приводит к следующему рекуррентному уравнению:

$$\mathbf{y}_i(l+1) = \mathbf{y}_i(l) - \alpha\frac{\partial E}{\partial \mathbf{y}_j} = \mathbf{y}_i(l) +$$

$$+\left[2\alpha\Big/\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\delta_{ij}\right)\right]\sum_{\substack{j=1\\ j\neq i}}^{N}\left[(\delta_{ij}-d_{ij})/(\delta_{ij}d_{ij})\right]\cdot\left[\mathbf{y}_i(l)-\mathbf{y}_j(l)\right]. \qquad (2)$$

Здесь α — настраиваемый параметр, l — номер итерации. Вычисление соотношения (2) прекращается в момент выполнения условия

$$E(l+1)-E(l)<\varepsilon\ , \qquad (3)$$

где $\varepsilon>0$ — некоторая константа. На практике значения ε обычно устанавливаются в интервале $10^{-4}...10^{-19}$.

К сожалению, описанный выше метод не позволяет добавлять во множество входных данных новые вектора без полного пересчёта всей карты. Одним из решений этой проблемы является использование триангуляции [6]. Суть этого подхода заключается в том, что в исходном пространстве находятся два ближайших к новому вектору узла. Зная расстояния до этих узлов, можно найти такое отображение, которое в точности сохраняло бы эти расстояния. Если существует более чем одна точка, удовлетворяющая данным условиям, то используется третий узел, для того чтобы принять решение о том, какую именно точку выбрать.

Другой подход к преодолению рассматриваемой проблемы состоит в использовании искусственной нейронной сети (ИНС) для интерполяции и экстраполяции отображения. В качестве ИНС можно использовать обычную нейронную сеть прямого распространения, обучаемую по методу обратного распространения ошибки. Однако её основным недостатком является большая вычислительная сложность. В качестве альтернативы в работе [7] была предложена ИНС, названная авторами SAMANN. Обучение данной ИНС происходит без учителя, что позволяет уменьшить вычислительные затраты. Основным недостатком ИНС является требование большого объёма обучающей выборки, который не всегда можно обеспечить.

Следует также учитывать, что использование триангуляции и ИНС не позволяет получить адекватного отображения в том случае, если поступающие входные данные довольно сильно отличаются от обучающего множества. В связи с этим в данной статье предлагается модификация метода Сэммона на случай добавления в отображаемое множество X ещё одного элемента.

Модифицированный алгоритм. Обозначим через Δ матрицу расстояний между векторами множества X:

$$\Delta=\begin{bmatrix}\delta_{11} & \delta_{12} & \dots & \delta_{1N}\\ \delta_{21} & \delta_{22} & & \delta_{2N}\\ \vdots & & \ddots & \vdots\\ \delta_{N1} & \delta_{N2} & \dots & \delta_{NN}\end{bmatrix}. \qquad (4)$$

После выполнения отображения в пространство размерности q по алгоритму (1)–(3) мы получим множество векторов Y, каждый элемент которого содержит в себе координаты некоторой точки в этом пространстве. Расстояния между точками в пространстве отображения при этом будет определяться так:

$$\boldsymbol{D}=\begin{bmatrix}d_{11} & d_{12} & \dots & d_{1N}\\ d_{21} & d_{22} & & d_{2N}\\ \vdots & & \ddots & \vdots\\ d_{N1} & d_{N2} & \dots & d_{NN}\end{bmatrix}. \qquad (5)$$

Тогда при добавлении ещё одного вектора x_{N+1} матрица расстояний Δ будет выглядеть следующим образом:

$$\Delta^* = \left[\begin{array}{cccc|c} \delta_{11} & \delta_{12} & \cdots & \delta_{1N} & \delta_{1N+1} \\ \delta_{21} & \delta_{22} & & \delta_{2N} & \delta_{2N+1} \\ \vdots & & \ddots & \vdots & \vdots \\ \delta_{N1} & \delta_{N2} & \cdots & \delta_{NN} & \delta_{NN+1} \\ \hline \delta_{N+11} & \delta_{N+12} & \cdots & \delta_{N+1N} & \delta_{N+1N+1} \end{array}\right]. \quad (6)$$

Матрица расстояний $\boldsymbol{D}^*$ в пространстве отображения определяется аналогично. В этом случае выражение (2) может быть переписано для расчёта только $N+1$ элемента отображения:

$$\mathbf{y}_{N+1}(l+1) = \mathbf{y}_{N+1}(l) + \left[2\alpha \Big/ \left(\sum_{i=1}^{N}\sum_{j=i+1}^{N+1}\delta_{ij}\right)\right] \sum_{\substack{j=1 \\ j\neq N+1}}^{N+1} \left[\left(\delta_{N+1j} - d_{N+1j}\right)/\left(\delta_{N+1j} d_{N+1j}\right)\right] \cdot \left[\mathbf{y}_{N+1}(l) - \mathbf{y}_j(l)\right] \quad (7)$$

Остановка работы алгоритма (7) происходит по выполнению условия (3).

Программа экспериментальных исследований. Для обучения были выбраны три фонемы английского языка, которые считаются наиболее трудными для произношения: /æ/, /ɔ/, /Λ/. Каждая фонема трижды проговаривалась группой из четырёх дикторов сначала в правильном варианте произношения, а затем в неправильном (т.е. звучание было близким к фонемам русского языка /э/, /о/, /а/). Полученные данные сначала записывались в память ПК в виде соответствующих звуковых файлов. Для этого применялась специальные программные и аппаратные средства: динамический микрофон AKG D77 S и ламповый микрофонный предусилитель ART TUBE MP Project Series USB. Частота дискретизации встроенного АЦП была установлена на уровне 8 кГц — общепринятая частота при обработке устной речи.

Реализация метода обучения правильному произношению может быть разделена на два независимых этапа: подготовительный и собственно этап обучения.

На подготовительном этапе формируется база априорных данных по каждой реализации всех трёх фонем. Для этого вычисляется вектор признаков, описывающий ту или иную реализацию. Для сравнения использовались четыре наиболее широко распространённых вида векторов признаков: коэффициенты линейного предсказания, кепстральные коэффициенты, коэффициенты линейного предсказания с деформированной частотной осью [8] и рассчитанные по ним кепстральные коэффициенты. В качестве меры расстояния между векторами была выбрана COSH-метрика при использовании обычных коэффициентов линейного предсказания и коэффициентов линейного предсказания с деформированной частотной осью. Данная метрика определяется так [9]:

$$d_{COSH}(S, S') = \frac{1}{2}\int_{-\pi}^{\pi}\left[\frac{S'(\omega)}{S(\omega)} + \frac{S(\omega)}{S'(\omega)}\right]\frac{d\omega}{2\pi} - 1. \quad (8)$$

Здесь $S(\omega)$ — СПМ эталонного сигнала, $S'(\omega)$ — СПМ входного сигнала. В качестве меры расстояния при использовании кепстральных коэффициентов и кепстральных коэффициентов линейного предсказания с деформированной частотной осью была выбрана евклидова

метрика. Для расчёта коэффициентов линейного предсказания использовался автокорреляционный метод, также известный как метод Дарбина [10]:

$$
\begin{aligned}
&r(m) = \sum_{n=0}^{N-1-m} x(n)x(n+m),\ 0 \le m \le p,\\
&E^{(0)} = r(0)\\
&k_i = \left\{ r(i) - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} r(|i-j|) \right\} \Big/ E^{(i-1)},\ 1 \le i \le p\\
&\alpha_i^{(i)} = k_i\\
&\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}\\
&E^{(i)} = (1 - k_i^2) E^{(i-1)}\\
&a_m = \alpha_m^{(p)},\ 1 \le m \le p
\end{aligned}
\qquad (9)
$$

Здесь r — вектор коэффициентов автокорреляции, p — порядок модели, a — вектор коэффициентов линейного предсказания, N — число отсчётов входного сигнала x.

Для вычисления коэффициентов линейного предсказания с деформированной частотной осью вектор коэффициентов автокорреляции пропускается через набор всепропускающих фильтров первого порядка следующего вида:

$$D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}\ . \qquad (10)$$

Здесь $-1<\lambda<1$ — коэффициент деформации. В том случае, когда $\lambda=1$, фильтр вырождается в обычную линию задержки. Параметр λ выбирается таким образом, чтобы получаемая частотная шкала была близка к шкале барк, и может быть приближенно рассчитан по следующей формуле:

$$\lambda \approx 1{,}0674 \left(\frac{2}{\pi} tan^{-1}(0{,}06583 f_s / 1000) \right)^{1/2} - 0{,}1916\,. \qquad (11)$$

Здесь f_s — частота дискретизации (Гц). В дальнейшем используется автокорреляционный метод расчёта коэффициентов линейного предсказания, однако коэффициенты автокорреляции для него вычисляются следующим образом:

$$
\begin{aligned}
&W^{(0)} = x\\
&r(0) = x^T x\\
&W^{(m)}(n) = -\lambda W^{(m-1)}(n) + W^{(m-1)}(n-1) + \lambda W^{(m)}(n-1),\ 1 \le n \le N\\
&r(m) = x^T W\\
&m = \overline{1, p}
\end{aligned}
\qquad (12)
$$

Кепстральные коэффициенты вычислялись по следующей формуле [11]:

$$
\begin{aligned}
&c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k}\ , \quad 1 \le m \le p\\
&c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k}\ , \quad m > p
\end{aligned}
\qquad (13)
$$

После вычисления набора векторов признаков производится их отображение на двумерную плоскость по алгоритму (1) — (3). Для определения качества такого отображения воспользуемся J-метрикой [12]. Эта метрика сравнивает рассеяние между классами с рассеянием внутри классов и определяется так:

$$J = tr\left(S_\omega^{-1} S_b\right) \to \max , \qquad (14)$$

где

$$\mathbf{S}_\omega = \sum_{i=1}^{M} \mathbf{P}(\omega_i)\Sigma_i ,$$

$$\mathbf{S}_b = \sum_{i=1}^{M} \mathbf{P}(\omega_i)\cdot(M_i - M_0)\cdot(M_i - M_0)^T ,$$

$$M_0 = \sum_{i=1}^{M} \mathbf{P}(\omega_i) M_i . \qquad (15)$$

Здесь Σ_i — корреляционная матрица i-го класса, M_i — его математическое ожидание, а $\mathbf{P}(\omega_i)$ — вероятность появления вектора данных, принадлежащего i-му классу. Большая величина J-метрики говорит о большей разделимости классов.

На этапе обучения диктор последовательно произносит различные реализации определённой фонемы. По каждой такой реализации вычисляется вектор признаков, который затем отображается на двумерную плоскость при помощи алгоритма (6) — (7). Таким образом, на плоскости, кроме множеств точек, соответствующих правильному и неправильному произношению фонемы, появляется ещё одна точка, которая характеризует расположение входного сигнала относительно данных множеств. Учитывая это, диктор может корректировать своё собственное произношение, приближая его к эталонному.

Результаты экспериментальных исследований. В качестве параметров алгоритма визуализации были выбраны следующие:

- порядок АР-модели $p=8$;
- число кепстральных коэффициентов $m=12$;
- параметр алгоритма отображения $\alpha=0{,}4$;
- относительная ошибка отображения $\varepsilon=10^{-5}$.

Начальные приближения y_0 задавались случайным образом в интервале от 0 до 1.

В таблице 1 показаны значения критерия (14) результатов отображения исходных данных на двумерную плоскость, полученные усреднением результатов для 10 различных начальных приближений .

Таблица 1

Фонема	Вектор признаков			
	Коэффициенты линейного предсказания	Кепстральные коэффициенты	Коэффициенты линейного предсказания с деформированной частотной осью	Кепстральные коэффициенты линейного предсказания с деформированной частотной осью
/æ/	22,24	11,85	49,63	21,5
/ɔ/	37,22	14,97	55,26	19
/Λ/	5,09	2,92	6,35	3,48

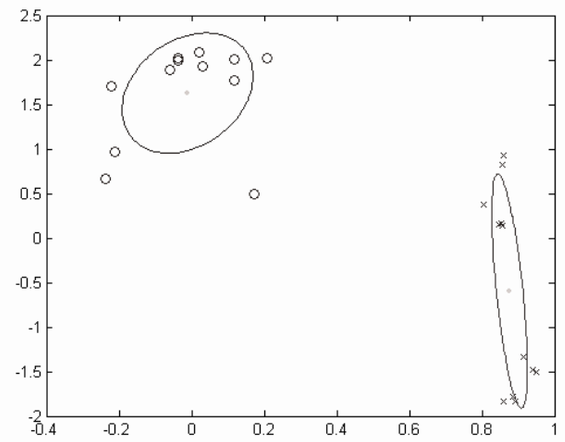

Рис. 1. Отображение реализаций фонемы /æ/ при использовании коэффициентов линейного предсказания с деформированной частотной осью

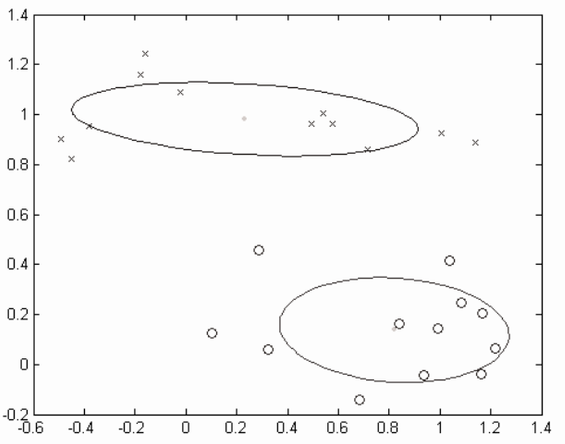

Рис. 2. Отображение реализаций фонемы /æ/ при использовании кепстральных коэффициентов линейного предсказания с деформированной частотной осью

Видно, что отображения классов фонем обладают наилучшей разделимостью при использовании коэффициентов линейного предсказания с деформированной частотной осью в качестве вектора признаков. Для наглядности получаемое для фонемы /æ/ отображение приведено на рисунке 1.

Здесь символами 'x' обозначены реализации, соответствующие правильному произношению, а символами 'o' — неправильному. Эллипсами показаны области, относящиеся к двум разным вариантам произношения данной фонемы. Несмотря на хорошую разделимость, видно, что отображаемые векторы имеют сильную корреляцию внутри кластера. Использование кепстральных коэффициентов линейного предсказания с деформированной частотной осью даёт худшие результаты, однако они менее коррелированны, чем при использовании коэффициентов линейного предсказания (в т.ч. коэффициентов линейного предсказания с деформированной частотной осью). На рисунке 2 приведён результат отображения для кепстральных коэффициентов линейного предсказания с деформированной частотной осью.

Меньшая коррелированность векторов признаков внутри класса может в перспективе облегчить процесс обучения правильному произношению. Кроме того, скорость сходимости алгоритма отображения при использовании кепстральных коэффициентов линейного предсказания с деформированной частотной осью примерно в два раза выше, чем при использовании в качестве признаков коэффициентов линейного предсказания с деформированной частотной осью (например, для обучающего множества фонемы /æ/ это около 550–650 итераций против более чем 1300).

В режиме обучения в качестве входных сигналов использовались произнесённые диктором различные реализации фонемы /æ/. На рисунке 3 показан результат работы системы. Символами «+» показано положение реализаций входных сигналов, произнесённых обучающимся диктором.

Из рисунка видно, что чем более чётко проговаривается звук /æ/, тем ближе к эталонному сигналу и дальше от области неправильного произношения располагаются реализации. Для остальных фонем были получены аналогичные результаты.

На рисунке 4 показан случай, когда произношение диктора резко отличается как от правильного, так и от неправильного вариантов.

Видно, что расположение точки, соответствующей входному сигналу, находится далеко за пределами обозначеных областей и позволяет адекватно оценивать данную ситуацию.

Выводы. Предложенный выше метод позволяет проводить обучение иностранным языкам и постановку произношения в режиме самостоятельного обучения. Показано, что использование коэффициентов линейного предсказания с деформированной частотной осью позволяет значительно улучшить разделимость кластеров, соответствующих различным вариантам произношения фонем. Вместе с тем, кепстральные коэффициенты линейного предсказания с деформированной частотной осью также могут использоваться в качестве векторов признаков. Основное их преимущество состоит в том, что, несмотря на худшую разделимость, отображаемые кластеры менее коррелированны. Это, в свою очередь, позволяет получать более наглядные результаты.

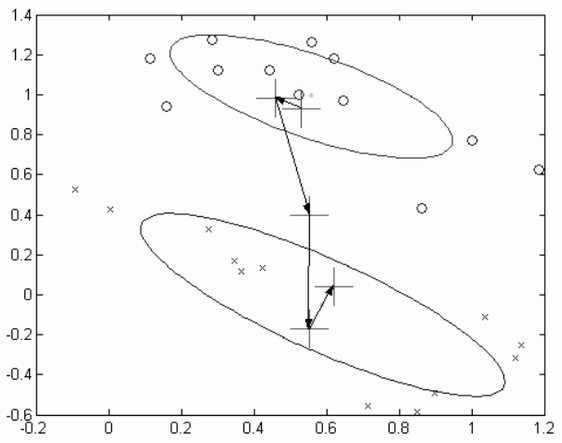

Рис. 3. Результаты обучения диктора

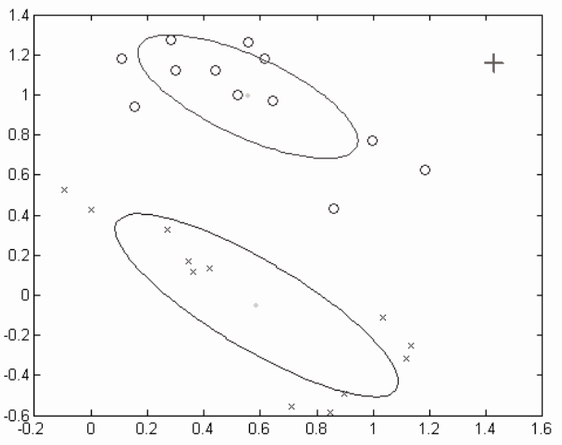

Рис. 4. Результат отображения для случая, когда произношение сильно отличается от эталонного

Дальнейшее развитие метода обеспечит возможность обучения произношению звуков внутри коротких слов. Выделение необходимых участков речевого сигнала при этом может осуществляться с помощью какого-либо алгоритма автоматического сегментирования. Наиболее перспективными здесь можно считать алгоритмы, основанные на аппарате скрытых марковских моделей [11]. Полученные результаты также могут найти применение и при обучении речи глухих и слабослышащих.

Литература

1. Hartis A. Computer-Based Audio-Visual Feedback Using Interactive Visual Displays for Speech Training. — PhD thesis. Department of Computer Science, University of Sheffield, 1999.

2. S. A. Zahorian and M. A. Zimmer, «Discriminative and Maximum Likelihood Classifiers for Computer-Based Visual Feedback and Speech Training for the Hearing Impaired,» World Multiconference on Systemics, Cybernetics and Informatics (SCI 2000/ISAS 2000), Vol. VI, Part II, pp. 475–479, Orlando, Florida, July 2000.

3. Borg I., Groenen P. Modern multidimensional scaling. Theory and applications. Spinger: 2005.

4. Demartines P., Herault J. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. IEEE Transactions on neural networks, 8(1):148–154, January 1997.

5. Sammon, J. W. A non-linear mapping algorithm for data structure analysis. IEEE Trans. Computers, CC.18(5):401–409. 1969.

6. R.C.T. Lee, J.R. Slagle, and H. Blum. A triangulation method for the dequential mapping of points from nspace to twospace. IEEE Transactions on Computers, C.26:288–292, 1977.

7. Mao J., Jain A.K. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*. 6(2):296–317. 1995.

8. Harma A. Frequency-warped autoregressive modeling and filtering. Dissertation for the degree of Doctor of Science in Technology. Department of Electrical and Communications Engineering, Helsinki University of Technology, Espoo, Finland, 2001.

9. A.H. Gray, Jr., and J.D. Markel, «Distance measures for speech processing» IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, pp. 380–391, Oct. 1976.

10. Марпл С.Л.-мл. Цифровой спектральный анализ и его приложения. М.: Мир, 1990.

11. Rabiner L.R., Juang B.-H. Fundamentals of speech recognition. Prentice Hall, Englewood Cliffs, NJ, 1993.

12. Friedman H.P., Rubin J. On some invariant criteria for grouping data. — Amer. Stat. Assoc. J. 62, 1159–1178. 1967.

13. Herault J., Guerin-Dugue A., Villemain P. Searching for the embedded manifolds in highdimensional data, problems and unsolved questions. ESANN'2002 proceedings 24-26 April 2002. Pp. 173–184.

14. *Фукунага К.* Введение в статистическую теорию распознавания образов. / Пер. с англ. М.: Наука. 1979.

Губочкин Иван Вадимович —

инженер-программист ООО «МФИ-Софт», аспирант кафедры математики и информатики Нижегородского государственного лингвистического университета. Область научных интересов — автоматическая обработка речевых сигналов. Автор шести научных работ.