НАУЧНОЕ ИЗДАНИЕ МГТУ ИМ. Н. Э. БАУМАНА
НАУКА и ОБРАЗОВАНИЕ
Эл № ФС77 - 48211. Государственная регистрация №0421200025. ISSN 1994-0408
электронный научно-технический журнал

Адаптивная кластерная модель минимальных речевых единиц в задачах анализа и распознавания речи

 # 02, февраль 2013
DOI: 10.7463/0213.0527867
Савченко В. В., Акатьев Д. Ю.
УДК 004.934

Россия, Нижегородский государственный лингвистический университет им. Н.А. Добролюбова
svv@lunn.ru
akatjev@lunn.ru

Введение. При анализе устного текста на русском языке мы опираемся на наши точные знания в отношении его фонетического строя, количественного и качественного состава используемой фонетической системы, а также закономерностей ее функционирования в разговорной речи. Этими знаниями мы пользуемся, например, при транскрибировании потока речи. Однако если мы анализируем звучащий текст на неизвестном языке и нам недоступна информация, относящаяся к его тонкой структуре, то мы можем, либо, опираясь на наш лингвистический опыт, давать участкам речевого потока приблизительную интерпретацию в рамках Международного фонетического алфавита, либо, обратившись к акустическим понятиям, членить речь на некие минимальные звуковые единицы (МЗЕ) и давать им определенные метки. Очевидно, что второй подход со всех точек зрения наиболее информативен и универсален. Множество меток всех МЗЕ и составит, в таком случае, звуковой строй данного диалекта или языка.

Проблема состоит в том, что разговорная речь по своим акустическим характеристикам широко варьируется, причем не регулярным образом, не только от одного языка к другому, но и от одного носителя к другому носителю одного и того же языка. В указанных условиях становится проблематичной сама идея выделения *повторяющегося* набора МЗЕ из разговорного потока. Кроме того, длительность отдельных МЗЕ не превышает нескольких миллисекунд, и это главное препятствие для применения традиционных методов теоретической лингвистики к разговорной (устной) речи. С другой стороны, до настоящего времени проблема не была преодолена и методами экспериментальной фонетики. И главная причина здесь – отсутствие *адекватной системы описания отдельных фонем*.

В поисках путей решения указанной проблемы в недавно созданной информационной теории восприятия речи (ИТВР) [1] само понятие «фонема» впервые было строго определено в теоретико-информационном смысле как «*множество однородных МЗЕ, объединенных в кластер по критерию минимального информационного рассогласования (МИР) в метрике Кульбака-Лейблера»*. Условно говоря, человеческий мозг объединяет и запоминает в себе как нечто целое (в виде абстрактного образа) разные образцы (произношения) каждой отдельной фонемы в соответствующей «сфере» своей памяти вокруг абстрактного «центра» с заданным «радиусом» (рис. 1).

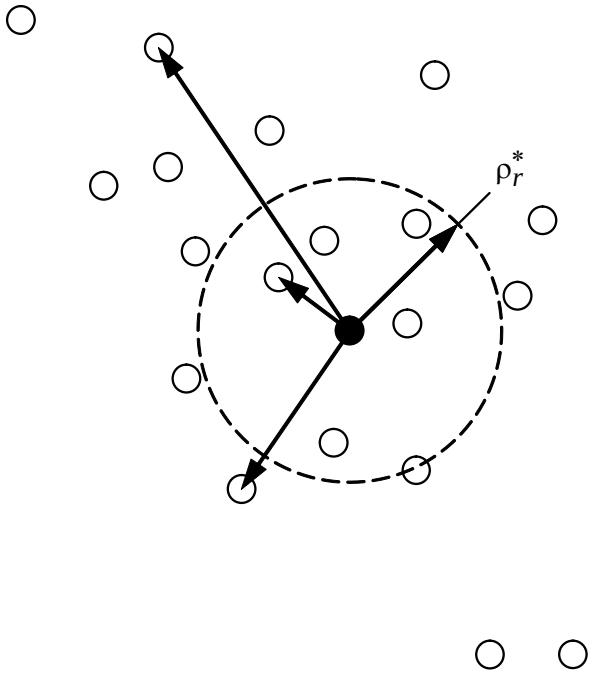

Рис. 1. Кластер реализаций фонемы и его информационный центр-эталон

Нетрудно понять, что этим определением одновременно решается множество актуальнейших проблем в области фонологического анализа: и вариативности разговорной речи, и априорной неопределенности, и адекватного описания звукового строя языка с кардинальным сжатием данных и, наконец, проблема обновления речевых баз данных (РБД) без разрушения их структуры.

Критерий МИР. Несмотря на существующие различия в реализациях некоторой r-ой фонемы все они воспринимаются человеком как нечто общее, иначе речь утратила бы свою информативность. Можно поэтому утверждать, что одноименные реализации $\mathbf{x}_{r,j}, \quad j=\overline{1,J_r}, \quad J_r >> 1$, в сознании человека группируются в соответствующие классы или речевые образы фонем $X_r = \{\mathbf{x}_{r,j}\}, \ r=\overline{1,R}$, вокруг некоторого центра – эталонной метки данного образа. В информационной теории восприятия речи указанные эталоны определяются в

строгом, теоретико-информационном смысле: речевая метка $\mathbf{x}_r^* \subset X_r$ образует *информационный центр-эталон* r-го речевого образа, если в пределах множества X_r она характеризуется минимальной суммой информационных рассогласований по Кульбаку-Лейблеру относительно всех других его меток-реализаций $\mathbf{x}_{r,j}$, $j = \overline{1, J_r}$.

Нетрудно увидеть, что именно в понятии информационного центра (ИЦ) r-го множества реализаций одноименных МЗЕ X_r дается наиболее информативное описание свойств соответствующей фонемы. Одновременно становится очевидным и механизм формирования самого этого множества. Сначала анализируемый (входной) речевой сигнал $X(t)$ в дискретном времени $t = 0,1,..$ разбивается на ряд последовательных сегментов данных $\mathbf{x}(t)$ длиной в одну МЗЕ: примерно 10–15 мс. После этого каждый такой парциальный сигнал рассматривается в пределах конечного списка фонем $\{X_r\}$ и отождествляется с той X_ν из них, которая отвечает критерию МИР относительно вектора $\mathbf{x}(t)$. Это известная формулировка критерия МИР в задачах автоматического распознавания речи. Задача существенно упрощается, если воспользоваться гауссовой (нормальной) аппроксимацией закона распределения каждой фонемы вида $\mathbf{P}_r = N(\mathbf{K}_r)$, где $\mathbf{K}_r$ - автокорреляционная матрица (АКМ) размера $n \times n$, $n \geq 1$.

Синтез адаптивного алгоритма. Предположим, что речевой образ каждой фонемы $X_r = \{\mathbf{x}_{r,j}\}$ представлен по-прежнему конечным (объема $J_r > 1$) множеством своих различных векторов-реализаций $\mathbf{x}_{r,j}$, $j = \overline{1, J_r}$, составленных из L последовательных во времени отсчетов одноименных МЗЕ $\{x_{r,j}(t)\}$ с периодом $T = 1/(2F) = const$. Здесь F - верхняя граница частотного диапазона речевого тракта. Рассматривая каждую такую реализацию в режиме «скользящего окна» длиной n отсчетов $(n << L)$, будем иметь $(L - n)$ векторов (столбцов) данных $\{\mathbf{x}_{r,j,i}\}$ размерностью $n = const$ каждый. Используя после этого формулу среднего арифметического, определим по ним выборочную оценку для АКМ гипотетического гауссова распределения $\mathbf{P}_r = N(\mathbf{K}_r)$.

Проблемы возникают, однако, в случае отсутствия априори классифицированных выборок $\{X_r\}$, т.е. при распознавании образов «без учителя». Автоматический анализ фонетического

состава речи чаще всего относится именно к такому кругу задач. И статистические характеристики фонем, и их используемое каждым диктором число R из общего списка зависят от особенностей его речевого аппарата. Здесь требуется алгоритм с самообучением, или адаптивный алгоритм фонетического анализа речи (ФАР). Для решения данной задачи в информационной теории разработан специальный инструмент: информационный $(R{+}1)$ -элемент. Информационный $(R{+}1)$-элемент – это условный термин, обозначающий устройство или алгоритм для автоматической классификации или распознавания сигнала $\mathbf{x}$ в пределах некоторого множества классов-альтернатив $\mathbf{P}_r$, $r = \overline{1,R}$. В основе его функционирования применяется статистический подход и критерий МИР. В отличие от других аналогичных алгоритмов с R выходами $(R{+}1)$ -элемент имеет дополнительный, $(R{+}1)$ -й выход, который сигнализирует об отказе при распознавании образов одновременно от всех R заданных альтернатив. Указанная особенность и служит основой для построения эффективного алгоритма распознавания образов в условиях априорной неопределенности. Задача сводится к последовательности задач статистической классификации «с учителем» при переменном (нарастающем) числе альтернатив R=1,2,…

Выделим в анализируемом речевом сигнале *X(t)* от некоторого диктора первые L отсчетов из соображений сохранения в них свойства приблизительной стационарности или однородности распределения $\mathbf{P}_r$. Например, при стандартной частоте дискретизации телефонного канала связи в 8 кГц обычно полагают $L = 100…200$ (это те же 10 – 15 мс). Используем полученный минимальный сегмент данных $\mathbf{x}_1 = \{x_1,...,x_L\}$ в качестве обучающей выборки X_1 для оценивания АКМ первой МЗЕ из сигнала. Соответствующий закон распределения $\mathbf{P}_1 = N\left(\hat{\mathbf{K}}_1\right)$ – это первый из элементов нашего будущего списка. После этого приравниваем $R = 1$ и берем второй сегмент выборки для анализа: $\mathbf{x}_2 = \{x_{L+1},...,x_{2L}\}$. Следуя выражению для решающей статистики МИР, определим для него удельную величину информационного рассогласования (ВИР) [2]

$$\rho\left(X_2, X_r\right) = \rho_r\left(\mathbf{x}\right)\Big|_{\mathbf{x}=\mathbf{x}_2} \tag{1}$$

относительно первой МЗЕ (при равенстве $r = 1$). Полученный результат сопоставляется с порогом по ВИР в роли допустимой величины рассогласований между разными реализациями одних и тех же фонем устной речи:

$$\rho\left(X_2, X_r\right) \le \rho_0. \tag{2}$$

При нарушении данного неравенства в нашем начальном списке фонем появится второй элемент, и вслед за этим приравниваем число выявленных фонем $R=2$. В противном случае принимается решение об объединении выборок X_1 и X_2 в один речевой образ $\mathbf{P}_1$: в качестве или одной МЗЕ удвоенной длительности $L_r = 2L$, если выборки смежные, или двух разных реализаций первой фонемы, если выборки не стыкуются. Равенство $R=1$ в обоих случаях сохраняется.

Нетрудно понять, что в форме условия (2) реализуется проверка гипотез об однородности выборок, а понятие фонемы определяется здесь как кластер однородных МЗЕ по критерию МИР. Это типичная формулировка информационного $(R+1)$-элемента.

Фонетический анализ речи. Вычисления по схеме (1), (2) повторяются циклически для всех последующих сегментов данных из речевого сигнала $X(t)$, причем повторятся «нарастающим итогом» для переменного значения $R=2,3,\ldots$ Каждый очередной сегмент данных сопоставляется по правилу (2) одновременно со всеми R множествами $\{X_r\}$ из текущего списка фонем. При этом не исключается возможность объединения одного и того же сегмента данных с элементами одновременно нескольких разных множеств. В результате будем иметь список фонем с некоторым фиксированным числом элементов R^*. Это важная характеристика как анализируемого речевого сигнала, так и самого диктора. Чем больше значение R^* для конкретного диктора, тем богаче с фундаментальной, фонетической точки зрения его речь. В данном выводе и состоит, по-видимому, главный смысл и назначение фонетического анализа речи (ФАР). Однако здесь же возникает и очевидная проблема: чрезмерно большое число фонем в речи диктора – это признак ее нечеткости, или не информативности. С точки зрения качества устной речи первостепенный интерес, безусловно, представляет собой множество четких МЗЕ. Его, в таком случае, и следует считать основным итогом ФАР. Поэтому логика подсказывает: после выполнения всех перечисленных выше вычислений некоторые «фонемы» из окончательного списка можно исключить как маргинальные.

Добавим к сказанному, что предложенный алгоритм имеет множество разнообразных модификаций за счет, главным образом, применения рекуррентных вычислительных процедур корреляционно-спектрального анализа. Среди них наибольший интерес представляет метод обеляющего фильтра (МОФ), основанный на авторегрессионной модели МЗЕ.

В ранних работах [1-3] было показано, что в асимптотике, когда $n \to \infty$, и при гауссовом распределении речевого сигнала $\mathbf{P}_r = N(\mathbf{K}_r)$ с обратной АКМ ленточной структуры выражение для оптимальной решающей статистики из выражения (1) сводится к виду

$$\rho_{x,r} = \frac{1}{F+1} \sum_{f=0}^{F} \frac{\left|1 + \sum_{m=1}^{p} a_r(m) e^{-j\pi mf/F}\right|^2}{\left|1 + \sum_{m=1}^{p} a_x(m) e^{-j\pi mf/F}\right|^2} - 1 \geq 0. \quad (3)$$

Здесь $\{a_x(m)\}$, $\{a_r(m)\}$ - два вектора авторегрессионных -коэффициентов: входного сигнала и *r*-го эталона, оба одного порядка $p>1$. Это стандартная формулировка МОФ в частотной области. Преимуществом данной интерпретации критерия МИР является, прежде всего, возможность его эффективной реализации в адаптивном варианте на основе быстрых вычислительных процедур авторегрессионного анализа, таких как метод Берга и др. Именно такой вариант МОФ был реализован в дальнейшем для проведения его экспериментальных исследований в типовой задаче ФАР.

Программа и результаты экспериментальных исследований. Для экспериментальных исследований предложенного алгоритма (1)...(3) была разработана информационная система фонетического анализа, обучения и тестирования слитной речи, основной интерфейс которой показан на рис. 2.

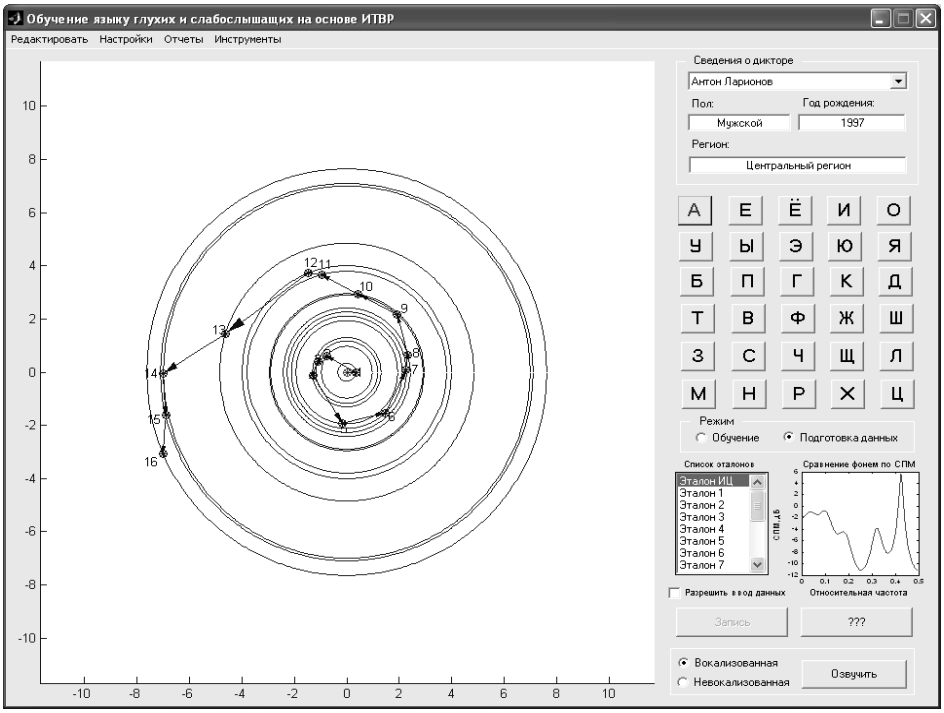

Рис. 2. Интерфейс информационной системы фонетического анализа, обучения и тестирования слитной речи

Программа экспериментальных исследований была разбита на два этапа [4]. На первом этапе осуществлялось формирование базы эталонов МЗЕ по группе тестируемых дикторов, а на втором – исследование особенностей звукового строя речи тех же дикторов в комфортных и некомфортных условиях. На обоих этапах для работы применялись 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 равной 8 кГц – общепринятая частота при обработке устной речи. Испытания проводились на ноутбуке следующей конфигурации: Asus X50V, 1024 Мбайт ОЗУ, Windows XP, Matlab 6.5. Формирование фонетической базы эталонов происходило следующим образом.

Вначале для каждой из основных (продолжительных) фонем русского языка было записано в комфортных условиях по одному образцу МЗЕ от выбранного диктора-мужчины. Затем к этим образцам были добавлены эталоны того же диктора в тех же условиях, но произнесённые в разное время суток. При этом диктор произносил каждую фонему по 15-20 раз. Звуковой сигнал вводился в информационную систему в реальном времени в режиме «Подготовка данных». Всего, таким образом, было сформировано шесть персональных баз эталонов от шести дикторов-мужчин, а также две базы эталонов от дикторов-женщин.

На втором этапе каждый диктор в заведомо менее комфортных условиях: в нашем случае – после значительной физической нагрузки (пульс 140-160 ударов в мин.) произносил каждую из 21 фонем по 10 – 15 раз. И каждый раз информационной системой фиксировался соответствующий результат: текущее значение ВИР по отношению к заранее сформированной базе эталонов. Цель данного эксперимента – выбрать из общего списка фонем национального языка те фонемы, которые наиболее остро реагируют в своих реализациях на условия произнесения их диктором. Смысл этой цели очевиден: настраивая информационную систему на наиболее чувствительные фонемы, мы гарантируем максимальную чувствительность нашего восприятия по отношению к эмоциональному и физическому состоянию диктора. Важнейший момент – это количественная характеристика степени возбуждения диктора, а именно: ВИР между фонемами в текущем сигнале и их эталонами. Для иллюстрации сказанного на рисунках ниже представлены две диаграммы ВИР при произнесении фонемы «Х» некоторым диктором-мужчиной в комфортных (рис. 3) и некомфортных (рис. 4).условиях. Здесь центр окружностей характеризует положение первого эталона в пределах Х-кластера одноименных МЗЕ.

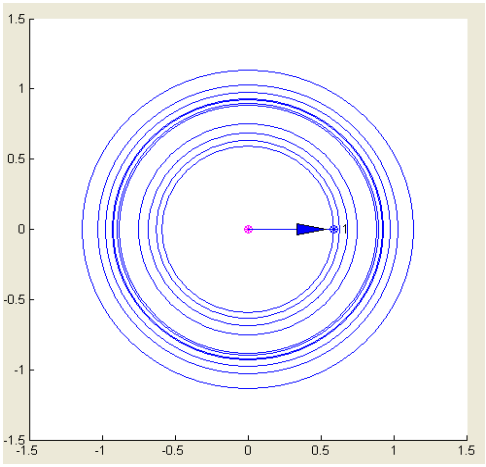

Рис. 3. Величина информационного рассогласования при произнесении фонемы «Х» диктором-мужчиной в *комфортных* условиях

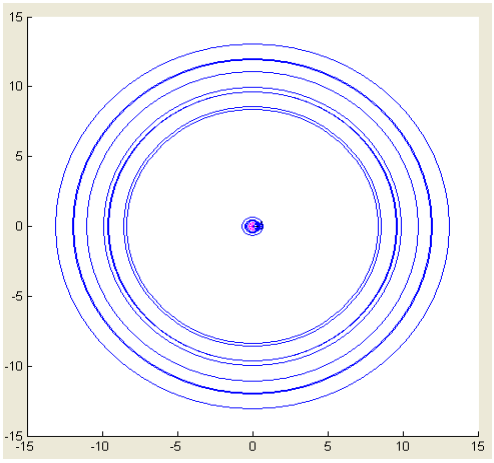

Рис. 4. Величина информационного рассогласования при произнесении фонемы «Х» диктором-мужчиной в *некомфортных* условиях

А каждая окружность – это результат очередного произнесения фонемы. Ее радиус определяется значением ВИР по отношению к эталону. Чем больше радиус, тем хуже качество произнесения. Видно, что при изменении условий на некомфортные в среднем на порядок (!) увеличилась вариативность произнесений данного диктора (см. шкалу делений по оси абсцисс). Аналогичные результаты были получены и для других дикторов из контрольной группы. Средние значения ВИР для типичных диктора-мужчины и диктора-женщины по всему списку фонем в зависимости от условий их произнесения представлены в следующей таблице.

Дикторы / **Фонемы**	**Диктор-муж.:** в комфортных условиях	в некомфортных условиях	**Диктор-жен.:** в комфортных условиях	в некомфортных условиях
А	0.77	0.53	1.02	3.84
О	7,6	20.03	6.4	16.4
У	3.12	4.06	12.7	8.3
Э	5.73	9.21	7.17	8.35
Ш	1.47	1.22	2.36	1.73
Щ	0.94	1.73	1.59	2.64
Р	0.72	2.49	2.03	2.71
С	0.58	0.51	0.82	0.93
В	1.38	1.14	1.82	1.92
З	3.87	4.69	4.14	4.73
Ж	0.69	0.71	0.94	1.19
И	2.51	3.04	3.92	4.08
М	1.94	6.14	1.01	3.32
Л	4.7	0.69	2.04	1.86
ЛЬ	2.19	1.54	1.91	1.79
Ф	1.78	1.91	1.83	1.89
Х	0.86	6.9	0.91	4.7
Ч	1.96	1.94	2.13	2.11
Е	3.81	4.57	5.17	5.89
Ы	2.49	3.18	3.67	4.29
Н	1.48	0.41	3.5	2.84

Здесь серым фоном отмечены наиболее чувствительные к условиям своего произнесения фонемы. Видно, что, по крайней мере, три из них: «Х», «М» и «О» одинаково высокочувствительны как в исполнении мужчин, так и женщин.

Заключение. К числу приоритетных направлений применения ИТВР и ее когнитивной кластерной модели МЗЕ (рис. 1) наряду с автоматической обработкой и распознаванием речи относятся, прежде всего, проблемы современной диалектологии. *Как сопоставить разные диалекты между собой по степени их объективной близости или различий на базовом, фонетическом, уровне? И какова количественная мера таких различий? Какие тенденции: сближения или удаления по фонетическому строю доминируют в настоящий момент в процессе исторического развития тех или иных диалектов? И, наконец, как можно лучше обучиться данному диалекту или, напротив, максимально ослабить его?* – Благодаря методологии ИТВР впервые в мировой науке открываются возможности дать четкие ответы на все перечисленные выше и подобные им вопросы. В их изучении и состоит главная цель предлагаемого исследовательского подхода. А ожидаемые по результатам исследований выводы и обобщения должны составить материал для подготовки к изданию первой фонологической карты России с многоуровневым членением языкового ареала на родственные диалекты при учете степени их звуковых различий, т.е. впервые в практике лингвистического картографирования – с указанием переходных диалектных зон. Осуществление предлагаемого проекта стимулирует, в свою очередь, научные исследования в области не только современной лингвистики, но и информатики в целом, прежде всего, прикладной информатики. Полученные результаты открывают качественно новые возможности для решения целого ряда актуальных задач, которые до настоящего времени остаются не решенными или решены неудовлетворительно, в том числе:

1) создание персональных (под каждого диктора) речевых баз данных;

2) анализ качества устной речи на базовом, фонетическом уровне;

3) автоматическое тестирование качества систем речевой связи

и другие.

Список литературы

1. Савченко В.В. Информационная теория восприятия речи // Изв. вузов России. Радиоэлектроника. 2007. Вып. 6. С. 3-9.

2. Савченко В.В., Акатьев Д.Ю. Технология обучения и тестирования речи на основе когнитивной кластерной модели минимальных речевых единиц // Нелинейная динамика в когнитивных исследованиях: сб. трудов Всерос. конф. Н. Новгород, 2011. С. 175-177.

3. Савченко В.В. Различение случайных сигналов в частотной области // Радиотехника и электроника. 1997. Т. 42, № 4. С. 426-431.

4. Савченко В.В. Автоматическое распознавание речи на основе кластерной модели минимальных речевых единиц в информационной метрике Кульбака-Лейблера. // Изв. вузов России. Радиоэлектроника. 2011. Вып. 3. С. 9-19.

SCIENTIFIC PERIODICAL OF THE BAUMAN MSTU
SCIENCE and EDUCATION
EL № FS77 - 48211. №0421200025. ISSN 1994-0408
electronic scientific and technical journal

Adaptive cluster model of minimal speech units in analysis and speech recognition problems

 # 02, February 2013
DOI: 10.7463/0213.0527867
Savchenko V.V., Akat'ev D.Yu.

Russia, Linguistics University of Nizhny Novgorod
svv@lunn.ru
akatjev@lunn.ru

This article deals with the problem of variability of word pronunciation in analysis and speech recognition tasks. An adaptive acoustic model defined as a multitude of minimal sound units (MSU) united into a cluster-phoneme under the principle of minimum informational mismatch in Kullback-Leibler metric, is proposed. An adaptive algorithm of filling the MSU cluster from a continuous stream of speech was developed on the basis of the whitening filter method. An example of its practical implementation is also provided in the article. As a result of this experiment, from the total list of phonemes of the national language the authors selected the phonemes which, in their implementation, are the most sensitive to conditions of their pronunciation by the speaker. Adjusting an information system to such a phoneme, the authors guarantee maximum sensitivity of perception in relation to the speaker's emotional and physical state.

Publications with keywords: automatic speech recognition, informative mismatch, adaptive cluster model, speech units
Publications with words: automatic speech recognition, informative mismatch, adaptive cluster model, speech units

References

1. Savchenko V.V. Informatsionnaia teoriia vospriiatiia rechi [The information theory of speech perception]. *Izv. vuzov Rossii. Radioelektronika*, 2007, no. 6, pp. 3-9.

2. Savchenko V.V., Akat'ev D.Iu. Tekhnologiia obucheniia i testirovaniia rechi na osnove kognitivnoi klasternoi modeli minimal'nykh rechevykh edinits [The technology of training and testing of speech on the basis of cognitive cluster model of minimal speech units]. *Nelineinaia dinamika v kognitivnykh issledovaniiakh: sb. trudov vseros. konf.* [Nonlinear dynamics in cognitive studies: proc. of all-Russian conf.]. Nizhny Novgorod, 2011, pp. 175-177.

3. Savchenko V.V. Razlichenie sluchainykh signalov v chastotnoi oblasti [The distinction between random signals in the frequency domain]. *Radiotekhnika i elektronika*, 1997, vol. 42, no. 4, pp. 426-431.

4. Savchenko V.V. Avtomaticheskoe raspoznavanie rechi na osnove klasternoi modeli minimal'nykh rechevykh edinits v informatsionnoi metrike Kul'baka-Leiblera [Automatic recognition of speech on the basis of cluster models of speech units in the Kullback-Leibler information metric]. *Izv. vuzov Rossii. Radioelektronika*, 2011, no. 3, pp. 9-19.