



Речевые

ТЕХНОЛОГИИ

3/2009

Главный редактор Александр Харламов

Состав редколлегии:

- Потапова Р.К., доктор филологических наук, профессор, заместитель главного редактора*
- Аграновский А.В., доктор технических наук, профессор*
- Женило В.Р., доктор технических наук*
- Жигулёвцев Ю.Н., кандидат технических наук*
- Кривнова О.Ф., доктор филологических наук*
- Кушнир А.М., кандидат психологических наук*
- Лобанов Б.М., доктор технических наук (Беларусь)*
- Максимов Е.М., доктор технических наук*
- Малеев О.Г., кандидат технических наук*
- Михайлов В.Г., доктор филологических наук*
- Нариньяни А.С., кандидат физико-математических наук*
- Петровский А.А., доктор технических наук (Беларусь)*
- Хитров М.В., кандидат технических наук*
- Чучупал В.Я., кандидат физико-математических наук*
- Шелепов В.Ю., доктор физико-математических наук (Украина)*
- Кушнир Д.А., ответственный секретарь, кандидат технических наук*

Содержание

Ермоленко Т.В., Фёдоров Е.Е.
Методика подавления цветных шумов в речевом сигнале3

Мурыгин К.В.
Распознавание визуальных частиц речи для обучения правильной артикуляции . . . 14

Людковик Т.В., Робейко В.В.
Озвучивание SMS-сообщений, отправляемых на стационарные телефоны24

Пилипенко В.В., Робейко В.В.
Опыт автоматического стенографирования украинской парламентской речи34



<i>Губочкин И.В.</i> Применение метода нелинейного отображения многомерных данных в задаче постановки правильного произношения звуков в составе отдельных слов	47
<i>Лукьяница А.А, Шишкин А.Г.</i> Автоматическое определение изменений эмоционального состояния по речевому сигналу	60
<i>Риехакайнен Е.И.</i> Перцептивно значимые элементы редуцированных словоформ	77
<i>Потапова Р.К.</i> Основные тенденции развития многоязычной корпусной лингвистики (часть вторая)	93

Редакция:

Редактор — Артём Ганькин
Корректор — Ирина Дёмина
Дизайн — Анна Ладанюк
Вёрстка — Сергей Бурукин

Адрес редакции: 109341, Москва, ул. Люблинская, д. 157, корп. 2.
Тел.: 8 (495) 979-54-27

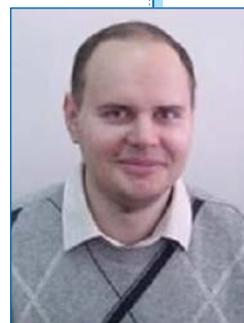
Подписано в печать 19.04.2010. Формат 60×90%. Бумага офсетная. Печать офсетная.
Печ. л. 14. Заказ № 0426. Издательский дом «Народное образование».
Отпечатано в типографии НИИ школьных технологий. 143500, г. Истра-2, ул. Заводская, д. 2А.
Тел.: 8 (901) 519-53-96, (495) 792-59-62.

© «Народное образование»

Методика подавления цветных шумов в речевом сигнале

Ермоленко Т.В.,
кандидат технических наук

Фёдоров Е.Е.,
кандидат технических наук



В задачах, связанных с распознаванием речи и идентификацией диктора, важную роль играет предварительная очистка сигнала от шума. В статье предлагается модификация известного метода спектрального вычитания – метода максимального правдоподобия, а также проводится численное сравнение реализованной методики со стандартными пакетами и методами, осуществляющими шумоочистку. Предложенный подход подавления шума в речевом сигнале основан на непрерывном вейвлет-преобразовании, продукционных правилах и учитывает акустические особенности фонетических классов звуков речи.

Noise clipping plays significant role in the problems of speech recognition and speaker identification. A modification of spectral subtraction method based on maximum likelihood estimation is proposed in this article. The numerical comparison of the realized technique, standard packages, and noise suppression methods is carried out. The offered approach of noise suppression in speech signal is based on continuous wavelet transformation, production rules and acoustic features of wide phonetic classes of speech sounds.

Введение

Анализ исследований. Для борьбы с помехами, рассредоточенными по спектру и пересекающимися с областью спектра речи, применяют методы спектрального вычитания [1–3] и пороговую вейвлет-обработку [4, 5].

Методы спектрального вычитания используют преобразование Фурье, их суть состоит в следующем. Вычисляется спектр шума $Y(k)$, после чего зашумлённый сигнал разбивают на фреймы, на каждом p -ом фрейме вычисляется спектр $X_p(k)$, из которого удаляется спектр шума. По полученному спектру $S_p(k)$ сигнал восстанавливается с помощью обратного преобразования:



$$S_p(k) = \begin{cases} H_p(k)X_p(k), & |X_p(k)|^v - \alpha |Y(k)|^v > 0, \\ \beta X_p(k), & \text{иначе} \end{cases}$$

где параметры α , β , v и функция $H_p(k)$ зависят от метода фильтрации.

Различают следующие методы: фильтрация по Берути, Шварцу и Макхоулу, основанная на вычитании энергетического спектра шума; винеровская фильтрация; метод максимального правдоподобия (MLEE); метод, использующий сглаживающий фильтр; метод EVRC (Enhanced Variable Rate Coder), реализованный в детекторе речи для мобильных телефонов; фильтрация по Болу, основанная на вычитании модуля амплитуды шума; метод, созданный на основе правила Эфраима и Малаха (EMSR).

В работе [1] были исследованы различные методы спектрального вычитания на белом и коричневом шуме. Стабильно хорошие результаты в обоих случаях показал метод MLEE.

Методы обработки сигналов, основанные на теории вейвлет-преобразований, лучше адаптированы к локальным свойствам сигнала и, в отличие от оконного преобразования Фурье, обеспечивают подвижное частотно-временное окно, которое сужается при высокой центральной частоте и расширяется при низкой. Теоретически доказано, что на всём множестве допустимых временных оконных функций спектрального оценивания вейвлет обладает лучшей разрешающей способностью по частоте и по времени [6, 7].

Традиционно методы пороговой вейвлет-обработки осуществляют подавление шумовой составляющей в сигнале с помощью быстрого вейвлет-преобразования и порогов [4, 5, 8]. Эти методы реализованы в пакете программ «Wavelet Toolbox» системы Matlab [4, 5]. Процесс понижения уровня шумов состоит из трёх этапов и включает в себя: разложение по вейвлет-базису, преобразование коэффициентов разложения, восстановление сигнала по преобразованным коэффициентам. На этапе преобразования детализирующие коэффициенты, не превышающие порог по абсолютному значению, обнуляются. Остальные детализирующие коэффициенты при использовании мягкого порога уменьшаются по модулю на его значение, при использовании жёсткого порога – остаются неизменными.

Однако быстрое вейвлет-преобразование имеет ряд ограничений [6, 8], которые снимаются при использовании более информативного непрерывного вейвлет-преобразования. Кроме того, все вышеперечисленные методы понижения уровня шума не учитывают фонетическую классификацию звуков речи, хотя на этапе предварительной обработки это могло бы понизить ошибки дальнейшего распознавания и повысить разборчивость речи.

Постановка задачи. На основе методов пороговой вейвлет-обработки разработать методику подавления цветного шума в речевом сигнале, учитывающую акустические особенности широких фонетических классов (ШФК) звуков речи.

Осуществить численное сравнение разработанной методики со стандартными пакетами и методами, осуществляющими шумоочистку.

1. Вейвлет-преобразование сигнала

Непрерывное вейвлет-преобразование одномерного сигнала $x(t)$ состоит в его разложении по базису, сконструированному из базового вейвлета $\psi(t)$ посредством масштабных изменений и смещений [9, 10], и может быть записано в виде:

$$CWT(a,b) = \int_{-\infty}^{\infty} x(t)|a|^{-1/2}\psi\left(\frac{t-b}{a}\right)dt. \quad (1)$$

Обратное вейвлет-преобразование имеет вид [6, 8]:

$$\tilde{x}(t) = \frac{1}{C_{\psi}} \int_{-\infty}^{\infty} \int_0^{\infty} CWT(a,b)|a|^{-1/2}\psi\left(\frac{t-b}{a}\right)\frac{dad b}{a^2}, \quad (2)$$

где C_{ψ} – нормирующий множитель.

Преобразование Фурье $\Psi(\omega)$ базового вейвлета является локализованной функцией с центром $\langle\omega\rangle$ и радиусом Δ_{ψ} , а вейвлет-преобразование для каждого значения масштабирующей переменной a в частотной области представляет собой полосовой фильтр с центральной частотой $\langle\omega\rangle/a$ и шириной полосы пропускания $2\Delta_{\psi}/a$ [10]:

$$win_a = [\langle\omega\rangle/a - \Delta_{\psi}/a; \langle\omega\rangle/a + \Delta_{\psi}/a]. \quad (3)$$

При обработке дискретных сигналов $x(n)$ конечной длины N возникает необходимость в использовании процедур квантования переменных a и b [10]:

$$a = a_0^i, \quad b = nb_0a_0^i, \quad i, n \in \mathbb{Z}, \quad 1 < a_0 < 2^{1/4}, \quad b_0 \neq 0, \quad (4)$$

а также в переходе к численному интегрированию в (1) и (2). Такой выбор шага изменения масштаба a_0 обусловлен исследованиями, проведёнными в [11], и предотвращает появление значительных колебаний на графике амплитудно-частотной характеристики фильтра.

В результате вейвлет-спектр и обратное вейвлет-преобразование вычисляются следующим образом:

$$d_{ml} = \sum_{n=0}^{N-1} x(n)a_0^{-m/2}\psi_{ml}(n)\Delta t, \quad l \in \overline{0, N-1}, \quad m \in \overline{i_{\min}, i_{\max}}, \quad (5)$$

$$\tilde{x}(n) = C_{\psi}^{-1} \sum_{m=i_{\min}}^{i_{\max}} \sum_{l=0}^{N-1} d_{ml}\psi_{ml}(n)\frac{\Delta a\Delta b}{a_0^{2m}} = C_{\psi}^{-1} \sum_{m=i_{\min}}^{i_{\max}} \sum_{l=0}^{N-1} d_{ml}\psi_{ml}(n)b_0(a_0 - 1), \quad (6)$$

где Δt – шаг квантования по времени, i_{\min}, i_{\max} – номер минимального и максимального уровня разложения,

$$\psi_{ml}(n) = a_0^{-m/2}\psi(a_0^{-m}n - b_0l).$$

Выбор начального (i_{\min}) и конечного (i_{\max}) уровней разложения, на которых проводятся обработка и анализ сигнала, осуществляется с учётом границ частотного диапазона сигнала $[f_1; f_2]$ (для речевых сигналов $f_1=100$ Гц; $f_2=8$ кГц), а также частоты дискретизации f_d .

Пользуясь (3)–(4), для банка вейвлет-фильтров можно получить значения границ полос пропускания на каждом уровне разложения m :



$$win_m = [\langle \omega \rangle / a_0^m - \Delta\Psi / a_0^m; \langle \omega \rangle / a_0^m + \Delta\Psi / a_0^m], i_{\min} \leq m \leq i_{\max}.$$

Таким образом, увеличение уровня разложения приводит к понижению центральной частоты соответствующего фильтра; следовательно, для анализа частотного диапазона $[f_1; f_2]$ с помощью банка вейвлет-фильтров необходимо, чтобы:

$$\frac{\langle \omega \rangle}{a_0^{i_{\min}}} = \frac{2\pi f_2}{f_d}, \quad \frac{\langle \omega \rangle}{a_0^{i_{\max}}} = \frac{2\pi f_1}{f_d},$$

откуда получаем границы диапазона уровней разложения:

$$i_{\min} = \text{int} \left(\log_{a_0} \frac{f_d \langle \omega \rangle}{2\pi f_2} \right), \quad i_{\max} = \text{int} \left(\log_{a_0} \frac{f_d \langle \omega \rangle}{2\pi f_1} \right),$$

где int – округление до ближайшего целого.

2. Подавление цветного шума в речевом сигнале

Методика подавления цветного шума в сигнале базируется на методах вейвлет-анализа, шумоочистки, продукционных правил, наряду с адаптацией к шуму учитывает акустические особенности ШФК звуков речи и состоит из четырёх этапов:

- вычисление порогов по образцу шума (обучение шуму);
- маркировка фреймов сигнала;
- удаление шума из вейвлет-образа сигнала;
- восстановление сигнала по обновлённым коэффициентам.

2.1. Обучение шуму

На этапе обучения шуму выполняется:

- 1) Вейвлет-разложение (5) сигнала $\varepsilon(n)$ длиной N_ε , содержащего образец шума, по уровням $i = \overline{i_{\min}; i_{\max}}$.

Затем этот сигнал разбивается на фреймы длиной ΔN отсчётов. В пределах одного фрейма сигнал является однородным. Для речевых сигналов длина фрейма $\Delta N \approx \Delta T f_d$, т.е. зависит от периода основного тона ΔT , который составляет не более 0,02с (50 Гц).

- 2) Для каждого s -го фрейма вычисляется мера контрастности (7):

$$C(m, s) = \lg \left(\frac{E_s(m)}{\sum_{j=j_{\min}}^m E_s(j)} \right), \quad m \in \overline{i_{\min} + 1; i_{\max}}, \quad (7)$$

где $E_s(m)$ – энергия вейвлет-спектра сигнала $\varepsilon(n)$ на уровне разложения m :

$$E_s(m) = \sum_{n=(s-1)\Delta N}^{s\Delta N} d_{mn}^2, \quad s \in \overline{1; \text{int}(N_\varepsilon / \Delta N)}. \quad (8)$$

Величина (7) характеризует распределение энергии спектра по уровням разложения и позволяет проводить анализ временной динамики энергии спектра сигнала. Операция логарифмирования даёт сглаженные характеристики, что предотвращает случайные всплески анализируемых величин в сигнале.

3) На каждом m -ом уровне разложения вычисляются несмещённые оценки математического ожидания $Aver(m)$ и дисперсии $D(m)$ величины $\{C(m, s)\}_{s=1}^{\text{int}(N_\varepsilon / \Delta N)}$, в результате чего получаются пороги

$$\alpha(m) = Aver(m) - \sqrt{D(m)}, \quad \beta(m) = Aver(m) + \sqrt{D(m)}$$

и усреднённые энергетические характеристики шума

$$E_\varepsilon(m) = \frac{\sum_{n=0}^{N_\varepsilon - 1} d_{mn}^2}{N_\varepsilon}. \quad (8)$$

2.2. Классификация фреймов

Классификация фреймов речевого сигнала $x(n)$, содержащего шум, выполняется на основе (7). При этом каждый s -ый фрейм может содержать один из четырёх ШФК звуков: только шум ($s \in Noise$); вокализованный звук ($s \in Voc$); шумный глухой щелевой или смычно-щелевой звук (Sh); шумный глухой смычный звук ($s \in P$).

Спектр вокализованных звуков, имеющих формантную структуру, сосредоточен, в основном, в частотной области от 100 Гц до 2.6 кГц [12].

Шумные глухие щелевые звуки и аффрикаты нестационарны, не имеют определённых спектральных параметров и характеризуются большой интенсивностью в диапазоне высоких частот от 4 кГц до 8 кГц в зависимости от звука и диктора [12–14].

Для проведения классификации выделяют два множества масштабов (уровней разложения):

— $M_{voc} = \{m : m_{voc} \leq m \leq i_{max}\}$ соответствует диапазону частот, где сосредоточена энергия вокализованных звуков (100–300 Гц), поэтому значения функции (7) при $m \in M_{voc}$ для фреймов, содержащих вокализованный звук, достаточно велики и превышают значения (7), полученные для фреймов сигнала $\varepsilon(n)$;

— $M_{sh} = \{m : i_{min} \leq m \leq m_{sh}\}$ соответствует высокочастотной части спектра (более 4 кГц), в которой сосредоточена энергия шумных глухих щелевых или смычно-щелевых звуков.

Для классификации каждого s -го фрейма сигнала $x(n)$ необходимы следующие действия.

1) Выполнить вейвлет-преобразование (5) сигнала $x(n)$.



2) Вычислить меру контрастности $C(m,s)$ для каждого s -го фрейма сигнала $x(n)$ согласно (7).

3) Отнести s -ый фрейм к одному из ШФК звуков по следующим правилам:

$$\forall m : \alpha(m) \leq C(m,s) \leq \beta(m) \rightarrow s \in \text{Noise} \vee P,$$

$$(\forall m \in M_{sh} : \alpha(m) > C(m,s)) \wedge (\exists n \in M_{voc} : C(n,s) > \beta(n)) \rightarrow s \in \text{Voc},$$

$$\exists m \in M_{sh} : \beta(m) < C(m,s) \rightarrow s \in \text{Sh},$$

а также получить массив маркировки фреймов $\{a_s\}_{s=1}^{N/\Delta N}$ на основе их классификации:

$$a_s = \begin{cases} 1, & s \in \text{Noise} \vee P \\ 2, & s \in \text{Sh} \\ 3, & s \in \text{Voc} \end{cases}$$

4) Для $i \in \overline{3, N/\Delta N - 4}$ провести корректировку маркировки фреймов:

4а) если один или два фрейма помечены неверно внутри одного звука:

$$a_{i-2} = a_{i-1} \wedge a_{i-1} \neq a_i \wedge a_{i-1} = a_{i+1} \wedge a_{i+1} = a_{i+2} \rightarrow a_i = a_{i+1}$$

(например, 11211→11111),

$$a_{i-2} = a_{i-1} \wedge a_{i-1} \neq a_i \wedge a_{i-1} \neq a_{i+1} \wedge a_{i-1} = a_{i+2} \wedge a_{i+2} = a_{i+3} \rightarrow \\ \rightarrow a_i = a_{i+2} \wedge a_{i+1} = a_{i+2}$$

(например, 112211→11111 или 112311→11111),

$$a_{i-2} = a_{i-1} \wedge a_{i-1} \neq a_i \wedge a_i \neq a_{i+1} \wedge a_{i+1} \neq a_{i+2} \wedge a_{i+2} \neq a_{i+3} \wedge a_{i+3} = a_{i+4} \rightarrow \\ \rightarrow a_i = a_{i+3} \wedge a_{i+1} = a_{i+3} \wedge a_{i+2} = a_{i+3}$$

(например, 1121211→111111, или 1121311→111111, или 1123211→111111);

4б) если один или два фрейма помечены неверно на границе звуков:

$$a_{i-3} = a_{i-2} \wedge a_{i-2} = a_{i-1} \wedge a_{i-1} \neq a_i \wedge a_i \neq a_{i+1} \wedge a_{i-1} \neq a_{i+1} \wedge a_{i+1} = a_{i+2} \wedge \\ \wedge a_{i+2} = a_{i+3} \rightarrow a_i = a_{i+1}$$

(например, 1112333→1111333),

$$a_{i-3} = a_{i-2} \wedge a_{i-2} = a_{i-1} \wedge a_{i-1} \neq a_i \wedge a_i \neq a_{i+2} \wedge a_{i-1} \neq a_{i+1} \wedge a_{i+1} \neq a_{i+2} \wedge \\ \wedge a_{i-1} \neq a_{i+2} \wedge a_{i+2} = a_{i+3} \wedge a_{i+3} = a_{i+4} \rightarrow a_i = a_{i-1} \wedge a_{i+1} = a_{i+2}$$

(например, 11122333→11113333),

$$a_{i-3} = a_{i-2} \wedge a_{i-2} = a_{i-1} \wedge a_{i-1} \neq a_i \wedge a_i = a_{i+2} \wedge a_{i-1} \neq a_{i+1} \wedge a_{i+1} \neq a_{i+2} \wedge \\ \wedge a_{i-1} \neq a_{i+2} \wedge a_{i+2} = a_{i+3} \wedge a_{i+3} = a_{i+4} \rightarrow a_{i+1} = a_{i+2}$$

(например, 11132333→11133333),

$$a_{i-3} = a_{i-2} \wedge a_{i-2} = a_{i-1} \wedge a_{i-1} \neq a_i \wedge a_i \neq a_{i+2} \wedge a_{i-1} = a_{i+1} \wedge a_{i+1} \neq a_{i+2} \wedge a_{i-1} \neq a_{i+2} \wedge a_{i+2} = a_{i+3} \wedge a_{i+3} = a_{i+4} \rightarrow a_i = a_{i-1}$$

(например, 11121333→11111333).

2.3. Удаление шума из вейвлет-образа сигнала

На этом этапе с использованием модифицированного метода MLEE и (8) производится изменение вейвлет-коэффициентов в соответствии с типом s -го фрейма. Возможны следующие варианты.

1) $a_s = 1$ (фрейм содержит шум):

$$\tilde{d}_{mn} = 0, \quad n \in \overline{(s-1)\Delta N, s\Delta N}, \quad m \in \overline{i_{\min} + 1; i_{\max}};$$

2) $a_s = 2$ (фрейм содержит шумные глухие щелевые и смычно-щелевые звуки):

$$\tilde{d}_{mn} = \begin{cases} \frac{1}{2} \left(1 + \sqrt{\frac{d_{mn}^2 - E_\varepsilon(m)}{d_{mn}^2}} \right) d_{mn}, & (d_{mn}^2 > E_\varepsilon(m)) \wedge m \in M_{sh}, \\ 0, & \text{иначе} \end{cases}$$

$$n \in \overline{(s-1)\Delta N, s\Delta N}, \quad m \in \overline{i_{\min} + 1; i_{\max}};$$

3) $a_s = 3$ (фрейм содержит вокализованные звуки):

$$\tilde{d}_{mn} = \begin{cases} \frac{1}{2} \left(1 + \sqrt{\frac{d_{mn}^2 - E_\varepsilon(m)}{d_{mn}^2}} \right) d_{mn}, & (d_{mn}^2 > E_\varepsilon(m)) \wedge m \in M_{voc} \\ 0, & \text{иначе} \end{cases},$$

$$n \in \overline{(s-1)\Delta N, s\Delta N}, \quad m \in \overline{i_{\min} + 1; i_{\max}}.$$

2.4. Восстановление сигнала

По обновлённым коэффициентам вейвлет-спектра с помощью обратного преобразования (6) сигнал восстанавливается:

$$\tilde{x}(n) = \sum_{m=i_{\min}}^{i_{\max}} \sum_{l=0}^{N-1} \tilde{d}_{ml} \psi_{ml}, \quad n \in \overline{0, N-1}.$$

Нормирующий множитель C_ψ вычисляется как отношение максимумов:

$$C_\psi = \frac{\max_n x(n)}{\max_n \tilde{x}(n)}.$$

Тогда окончательно имеем:

$$y(n) = C_\psi \tilde{x}(n), \quad n \in \overline{0, N-1}.$$



3. Численное исследование эффективности подавления шума с помощью разработанной методики

Для проведения численного исследования предложенная методика была реализована в программном комплексе.

Для предложенной методики использовался вейвлет Морле. В качестве сигнала было выбрано слово «Саша» с частотой дискретизации 22050 Гц, 8 бит, моно. Этот сигнал зашумлялся коричневым, розовым и белым шумом с помощью программы Adobe Audition.

На рис. 1–10 приведены: а) исходный сигнал (рис. 1), зашумлённый коричневым шумом (рис. 2); б) сигналы, очищенные с помощью аппроксимированного непрерывного вейвлет-преобразования (авторская методика) (рис. 3), стандартных методов спектрального вычитания и пакетов работы со звуком (рис. 4–10).

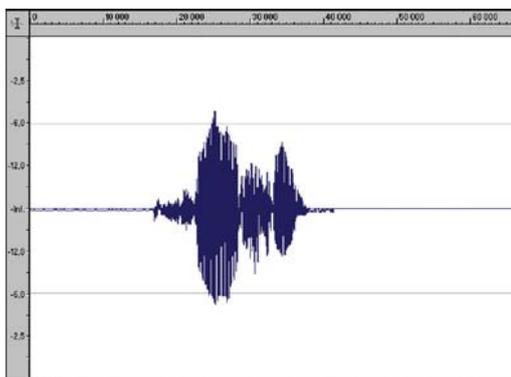


Рис. 1. Исходный сигнал

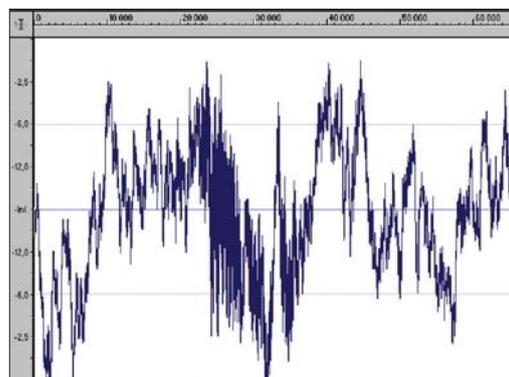


Рис. 2. Сигнал, зашумлённый коричневым шумом



Рис. 3. Сигнал, очищенный с помощью вейвлет-преобразования Морле

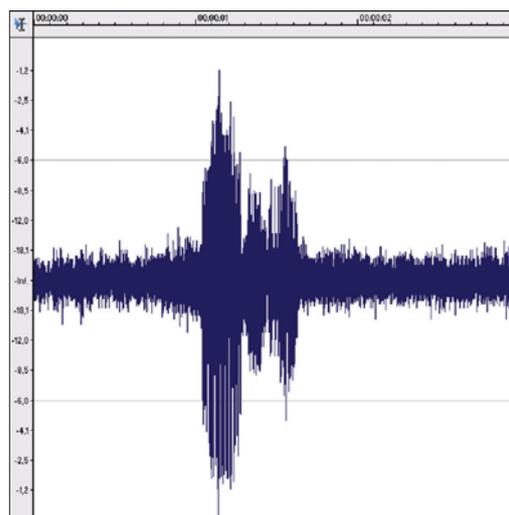


Рис. 4. Сигнал, очищенный с помощью MLEE

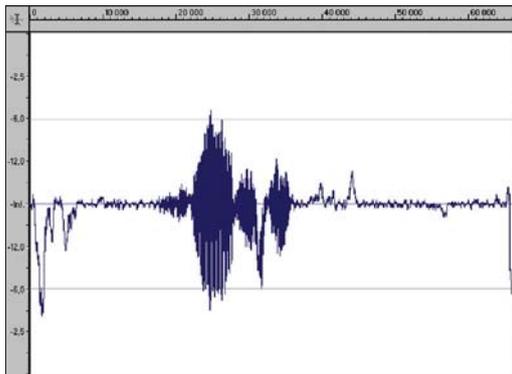


Рис. 5. Сигнал, очищенный с помощью пакета Clear Voice Denoiser

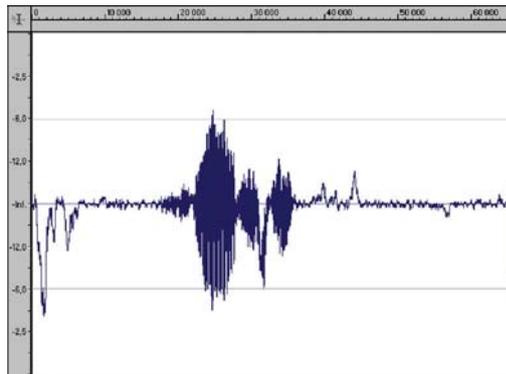


Рис. 6. Сигнал, очищенный с помощью пакета Adobe Audition v1

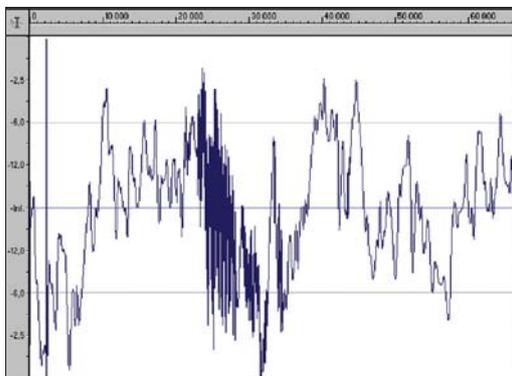


Рис. 7. Сигнал, очищенный с помощью жёсткого порога пакета Wavelets Extension Mathcad v11

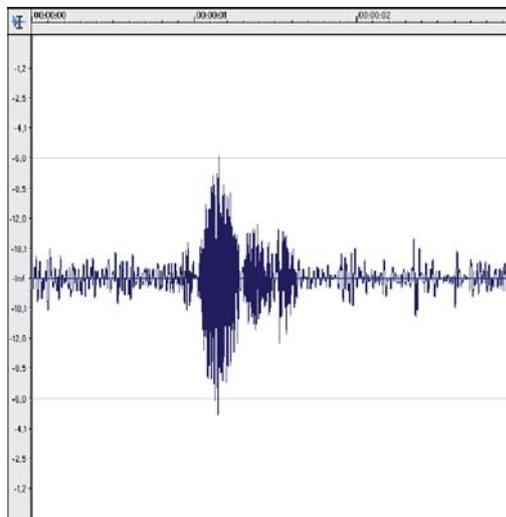


Рис. 8. Сигнал, очищенный с помощью фильтрации по Болю

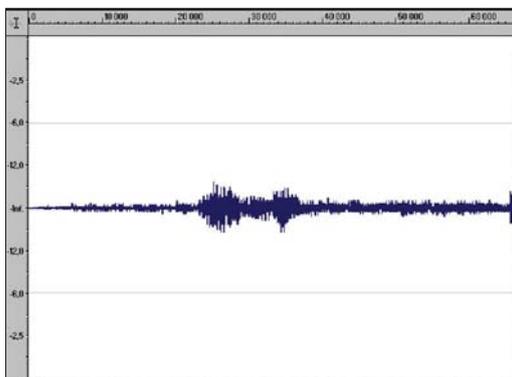


Рис. 9. Сигнал, очищенный с помощью EMSR

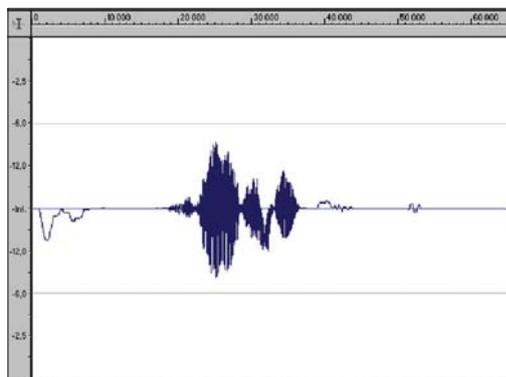


Рис. 10. Сигнал, очищенный с помощью пакета Nero WaveEditor v3

Чтобы провести оценку эффективности и сравнительный анализ результатов метода обесшумливания, реализованного в авторском программном комплексе, и методов обесшум-



ливания, реализованных в стандартных пакетах работы со звуком, авторами введена мера близости очищенного сигнала к исходному:

$$\delta = \sum_{s=1}^{N/\Delta N} \left(\frac{\sum_{n=0}^{\Delta N} (x_s^{source}((s-1)\Delta N + n) - x_s^{clear}((s-1)\Delta N + n))^2}{\sum_{n=0}^{\Delta N} (x_s^{source}((s-1)\Delta N + n))^2} \right),$$

где $x_s^{source}(n)$ – s -й фрейм исходного сигнала, $x_s^{clear}(n)$ – s -й фрейм очищенного сигнала, ΔN – длина фрейма.

Наилучшим будем считать метод, у которого $\delta \rightarrow \min$.

В таблице 1 приведены результаты численного исследования эффективности работы методов обесшумливания сигнала, зашумлённого цветными шумами. Как видно из таблицы 1, наилучшие результаты даёт авторская методика.

Таблица 1

Результаты численного исследования

Методики и пакеты	Значение δ		
	белый, отношение сигнал/шум 18дб	коричневый, отношение сигнал/шум 12дб	розовый, отношение сигнал/шум 25дб
Пакеты:			
Adobe Audition	12377	29419	24622
Clear Voice Denoiser	18122	356624	2840980
Nero WaveEditor	27442	14628	1342810
Wavelets Extension	17402	34514695	5918960
Методики:			
авторская	619796	12392900	6278350
одноканальный адаптивный фильтр	343049	114899000	9123650
Болла	41266	237512	387731
MLEE	59067	87095	180327
EMSR	11204	131935	202621
сглаживающий фильтр	113203	763875	8322940
EVRC	17402	34514700	5918960
жёсткий порог	619796	12392900	6278350

Заключение

Новизна. В статье была предложена модификация метода MLEE подавления цветного шума в сигнале, основанная на пороговой вейвлет-обработке и продукционных правилах. Благодаря учёту акустических особенностей ШФК звуков речи, подобный подход на этапе предварительной обработки позволяет понизить ошибки дальнейшего распознавания и повысить разборчивость речи.

Практическое значение. Основные результаты данной работы предназначены для реализации в системах распознавания речи и идентификации диктора, а также для создания интеллектуальных систем управления, в которых команды поступают на естественном языке.

Литература

1. *Фёдоров Е.Е.* Модели и методы преобразования речевых сигналов. Донецк: Норд-Пресс, 2006. 260 с.
2. *Recchione M.C.* The enhanced variable rate coder: Toll quality speech for CDMA // *International Journal of Speech Technology*. 1999. № 2. P. 305–315.
3. *Thiemann J.* Acoustic Noise Suppression for Speech Signals using Auditory Masking Effects. Montreal: McGill University, 2001. 83 p.
4. *Дьяконов В.П.* Вейвлеты. От теории к практике. М.: СОЛОН-Р, 2002. 448 с.
5. *Дьяконов В., Абраменкова И.* MATLAB. Обработка сигналов и изображений: Специальный справочник. СПб.: Питер, 2002. 608 с.
6. *Чуи К.* Введение в вейвлеты: Пер. с англ. М.: Мир, 2001. 412 с.
7. *Zhenilo V.R., Zhenilo M.V., Kalyuzhny D.N.* Fourier-Gauss Transform: Speech Signal Decomposition into Sonels. // *Proc. International Conf. on Speech and Computer (SPECOM'2007)*. Moscow (Russia). 2007. P. 259–29.
8. *Малла С.* Вейвлеты в обработке сигналов: Пер. с англ. М.: Мир, 2005. 671 с.
9. *Дремин И.М., Иванов О.В., Нечитайло В.А.* Вейвлеты и их использование // *Успехи физических наук*. 2001. Т. 171, №5. С. 465–501.
10. *Астафьева Н.М.* Вейвлет-анализ: основы теории и некоторые приложения // *Успехи физических наук*. 1998. №11. С. 1145–1170.
11. *Шитов А.Б.* Разработка численных методов и программ, связанных с применением вейвлет-анализа для моделирования и обработки экспериментальных данных: Дис. канд. техн. наук: 05.13.18. М., 2001. 125 с.
12. *Златоустова Л.В.* Фонетические единицы русской речи. М.: МГУ, 1981. 108 с.
13. *Фёдоров Е.Е., Шевцова И.А.* Численное исследование шипящих согласных звуков // *Искусственный интеллект*. 2004. №4. С. 661–665.
14. *Фёдоров Е.Е., Шевцова И.А.* Количественный анализ шумных глухих щелевых и смычно-щелевых звуков // *Искусственный интеллект*. 2005. №3. С. 308–313.

Ермоленко Т.В.

Кандидат технических наук, научный сотрудник отдела распознавания речевых образов Института проблем искусственного интеллекта МОН и НАН Украины. Распознаванием и обработкой речевых сигналов занимается с 2002 года.
etv@iai.donetsk.ua

Фёдоров Е.Е.

Доцент кафедры специализированных компьютерных систем Донецкой академии автомобильного транспорта.

В 2003 году защитил кандидатскую диссертацию в Институте проблем искусственного интеллекта МОН и НАН Украины. Автор более 75 научных публикаций, в том числе трёх монографий, посвящённых моделям и методам преобразования речевых сигналов. Основная область интересов: идентификация и верификация диктора, распознавание и синтез речи, анализ и синтез языковых конструкций, вибродиагностика и шумодиагностика, медицинская диагностика (анализ биосигнала).



Распознавание визуальных частиц речи для обучения правильной артикуляции

*Мурыгин К.В.,
кандидат технических наук*

В статье приводятся результаты исследований по проблеме распознавания визуальных частиц речи. Целью проведения исследований является разработка программной системы обучения правильной артикуляции при произнесении речи для упрощения её понимания слабослышащими людьми. Кроме этого, результаты исследований могут использоваться для дополнения звукового информационного канала визуальным, что необходимо для повышения качества распознавания речи в условиях шума или посторонних источников звука.

Results of researches devoted to recognition of visual speech particles are described in the article. The purpose of carrying out the researches is the development of program system for training of correct articulation during speech pronouncing for simplification of its understanding by deaf and hard of hearing people. Besides, the results of researches can be used for supplementing the sound information channel with the visual one, that is necessary for improvement of speech recognition quality in the conditions of noise or extraneous sound sources.

Введение

Большинство исследований в области распознавания речи ведутся в направлении интеллектуального анализа звуковой информации, которая считается наиболее информативной при передаче речевого кода. Этот факт проявляется, в частности, в значительных затруднениях, которые испытывает при общении человек с нарушениями слуха. Кроме того, фонетический состав языка является более полным, чем его визуальный алфавит. Так, в украинской речи встречаются 6 гласных и 32 согласные

фонемы, которые визуально можно представить не более чем 16 артикуляционными образами. Кроме восприятия звуковой речи, можно отметить существование возможности чтения с губ специально подготовленными людьми. Методики обучения этим навыкам, конечно, ориентированы, в основном, на слабослышащих и глухих и предполагают хорошее владение языком, на котором происходит разговор, его структурой, знание контекста, позволяющее получать дополнительную информацию на основе смыслового комбинирования. Автоматизация процесса зрительного восприятия речи влечёт необходимость применения достаточно качественного и сложного семантического анализатора, что предполагает использование не только математического аппарата распознавания зрительных образов.

Тем не менее, приведённые сложности решения общей задачи не исключают возможности получения важных, с практической точки зрения, результатов уже на начальных этапах, связанных с обнаружением и распознаванием области губ и её конфигурации. Таким результатом может быть система обучения правильной артикуляции для облегчения зрительного восприятия устной речи людьми с нарушением слуха. В настоящее время системы автоматического чтения с губ в большинстве своём разрабатываются для дополнения звукового информационного канала визуальным. В этой связи создание обучающей системы является новым направлением и имеет значительную практическую ценность.

Задача автоматического чтения по губам объединяет в себе несколько подзадач, некоторые из которых имеют самостоятельное практическое значение, а именно:

- обнаружение лиц (face detection);
- обнаружение или извлечение деталей лица, в частности области губ (face features extraction, lip tracking);
- выделение признаков для описания конфигурации губ, позволяющих находить соответствие с произнесённой фонемой, и разработка методов такой классификации (lip reading).

По аналогии с элементарными звуковыми частицами речи – фонемами – для обозначения элементарных визуальных образов речи будем использовать уже вполне устоявшийся термин – *визема*. Внимание данной статьи будет сконцентрировано на решении задачи классификации визем.

1. Формирование словаря визем

Согласно фонетике, в украинском языке существует 6 гласных и 32 согласные фонемы:

[i], [и], [e], [у], [o], [a];

[б], [п], [д], [д'], [т], [т'], [г], [к], [ф], [ж], [з], [з'], [ш], [с], [с'], [г], [х], [дж], [дз], [дз'], [ч], [ц], [ц'], [в], [й], [м], [н], [н'], [л], [л'], [р], [р'].

Здесь знак ' означает мягкость.

Сопоставляя фонетический состав украинского языка с исследованиями Бельтюкова В.И. для русского языка [1, 2] и учитывая фонетическое сходство украинского и русского языков, можно сформировать следующий визуальный алфавит украинских звуков (визем), представленный в табл. 1.



Таблица 1

Визуальный алфавит украинских звуков (визем), полученный по аналогии с алфавитом В.И. Бельтюкова

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
а	о	у	е	і	п	ф	ш	л'	р	с'	т	т'	к	й
				и	б	в	ж	р'		с	д	д'	г	
					м		ч			з'	н	н'	х	
							дж			з	л		г	
										ц				
										ц'				
										дз				
										дз'				

Предварительный анализ возможности автоматической классификации образов такого алфавита показал необходимость его существенного сокращения в направлении использования базовых, или опорных, визем [3]. В приведённой табл. 1 виземы, начиная с девятой, являются плохо различимыми даже человеком с его значительно более мощным зрительным аппаратом. Это во многом связано с тем, что процесс воспроизведения соответствующих им звуков в значительной мере скрыт внутри ротовой полости, что существенно усложняет их зрительное восприятие и, тем более, автоматическое распознавание на основе полученного цифрового изображения.

Поэтому в качестве рабочего алфавита визем принят алфавит, включающий в себя только опорные виземы (см. табл. 2).

Таблица 2

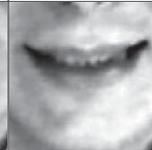
Принятый рабочий алфавит визем

1	2	3	4	5	6	7	8	9
а	о	у	е	і	п	ф	ш	§
				и	б	в	ж	
					м		ч	
							дж	

Для сравнения с более широким алфавитом (табл. 1) в таблице 3 приведены также изображения визем для элементов алфавита, не вошедших в рабочий алфавит (см. табл. 2).

Таблица 3

Элементы расширенного алфавита, не вошедшие в рабочий алфавит

9	10	11	12	13	14	15
л'	р	с'	т	т'	к	й
р'		с	д	д'	г	
		з'	н	н'	х	
		з	л		г	
		ц				
		ц'				
		дз				
		дз'				
						

Как видно из табл. 3, приведённые в ней элементы алфавита визуально трудно отличимы от элементов рабочего алфавита, что может существенно затруднить распознавание произнесённого звука по изображению соответствующей конфигурации губ. Так, виземы 10 и 12 визуально трудно отличимы от 8, а виземы 9, 11, 13 и 15 легко спутать как между собой, так и с виземой 5 принятого рабочего алфавита.

Знак \$ в рабочем алфавите визем означает нормальное положение, молчание, паузу или любую другую визему, не входящую в этот алфавит. Таким образом, при распознавании предпочтение отдаётся виземам 1–8. В случае отказа от распознавания (не распознана ни одна из восьми визем) данной конфигурации приписывается значение 9, которое также может генерироваться в промежуточном положении между двумя и более виземами.

2. Используемая база данных

Несмотря на то, что для создания описанной системы обучения необходимо решить несколько задач интеллектуальной обработки визуальной информации [4], будем считать, что положение области губ предварительно определено, например, методами [5, 6, 7]. Таким образом, на вход классификатора поступают вырезанные изображения области губ.

Для наполнения базы данных изображений визем была разработана специальная программа, позволяющая вырезать изображения визем из видеопотока. Пользователь сначала выбирает визему, которую он будет вводить, при этом ему демонстрируется пример правильного произнесения (см. рис. 1). После этого он самостоятельно воспроизводит выбранную визему, стараясь добиться максимального соответствия эталону, и сохраняет результирующее изображение на диск.

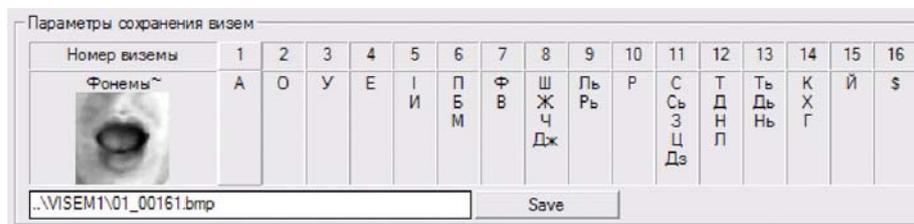


Рис. 1. Заполнение базы данных визем

С применением описанной программы была сформирована база изображений визем, произносимых разными людьми на разном расстоянии от камеры, насчитывающая 988 изображений разных размеров (см. рис. 2).



Рис. 2. Пример изображений визем из базы данных

Для исследований возможностей автоматического распознавания визем полученная база данных была размечена вручную. В ходе разметки на каждом изображении отмечались четыре точки, характеризующие крайние левое, правое, нижнее и верхнее положения точек губ. Для приведения изображений к одному масштабу за масштабный коэффициент был выбран горизонтальный размер области губ. Экспериментальная зависимость частот распределения отношения вертикального размера области губ к горизонтальному показана на гистограмме ниже (см. рис. 3).

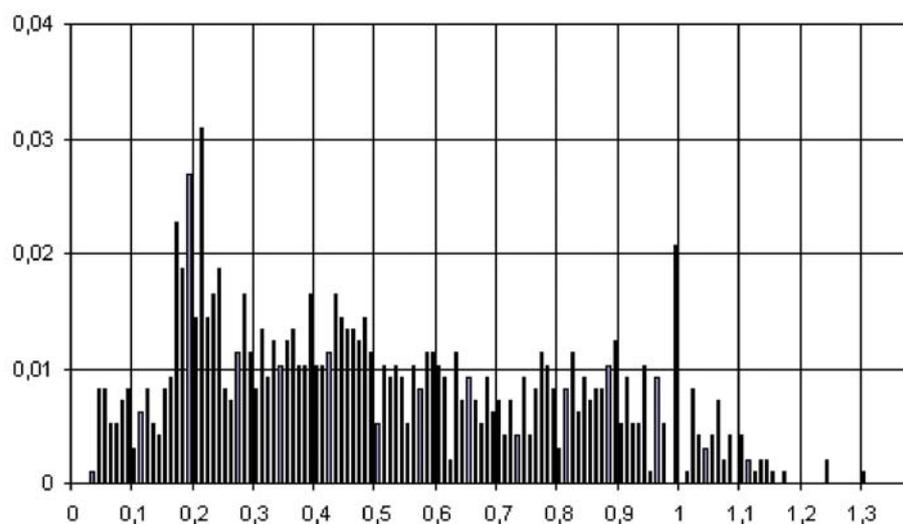


Рис. 3. Гистограмма частот отношения вертикального размера области губ к горизонтальному



Рис. 4. Пример изображений из базы данных области губ после уточнения их положения по имеющейся разметке

Из приведённой гистограммы следует, что основная часть изображений губ сосредоточена в интервале [0;1] для отношения вертикального размера области губ к горизонтальному. Это позволяет перейти к описанию области губ в виде квадрата, ширина и высота которого равны горизонтальному размеру губ, известному для каждого изображения из базы на основе сделанной разметки, с центром в точке, задаваемой выражениями:

$$X_u = \frac{X_{\max} + X_{\min}}{2}, Y_u = \frac{Y_{\max} + Y_{\min}}{2},$$

где X_{\max} , X_{\min} , Y_{\max} , Y_{\min} – соответственно крайние правые, левые, нижние и верхние координаты вручную размеченной области губ.

После обработки с учётом данных разметки на основе описанной методики получена база изображений, более точно описывающих область губ (см. рис. 4).

Полученная база допускает масштабирование с использованием стандартных методов изменения размеров изображений и удобна для использования при разработке и исследовании как методов обнаружения области губ, так и методов распознавания визем.

3. Используемые для классификации признаки

С точки зрения необходимости использования достаточно простых алгоритмов получения признаков, наиболее приемлемым является использование Хаар-подобных свойств, представляющих собой результат сравнения яркостей в двух прямоугольных областях изображения (см. рис. 5).

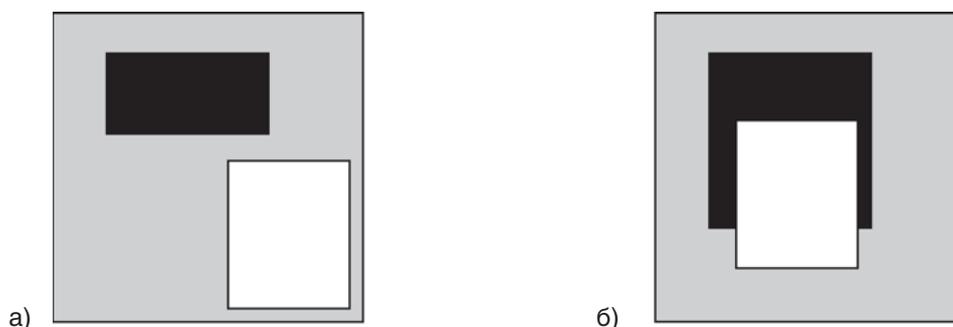


Рис. 5. Вид прямоугольных свойств, используемых в качестве признаков:
а) области не пересекаются; б) области пересекаются



Значение признака для данной области изображения или отклик области изображения на данное свойство вычисляется на основе следующего выражения:

в случае непересекающихся областей –

$$R = \begin{cases} 1, \frac{S_B}{N_B} > \frac{S_C}{N_C}; \\ -1, \frac{S_B}{N_B} \leq \frac{S_C}{N_C}. \end{cases}$$

в случае пересечения областей –

$$R = \begin{cases} 1, \frac{S_B}{N_B} > \frac{S_C - S_{C \cap B}}{N_C - N_{C \cap B}}; \\ -1, \frac{S_B}{N_B} \leq \frac{S_C - S_{C \cap B}}{N_C - N_{C \cap B}}. \end{cases}$$

Здесь индексы Ч и Б обозначают чёрную и белую области соответственно, а ЧПБ – область пересечения областей чёрного и белого цвета; S – сумма яркостей пикселей изображения, находящихся под областью; N – число пикселей изображения, находящихся под областью. Значения, получаемые на основе этих выражений, являются инвариантными по отношению к любым линейным преобразованиям функции яркости изображений, к которым с достаточной точностью можно отнести операции изменения яркости и контраста.

4. Обучение и тестирование классификаторов визем

Для решения задачи классификации конфигурации губ на входных изображениях, согласно введённому рабочему алфавиту визем, использовался подход, основанный на группировке описанных выше признаков в классификаторы с использованием алгоритма AdaBoost. Переход от задачи разделения двух классов был решён путём построения набора классификаторов, отделяющих каждую из визем принятого алфавита (см. табл. 2) от всех остальных визем. Такой подход позволяет получить больше информации о распознаваемом объекте за счёт возможности множественной классификации. При множественной классификации объект, поступивший на вход распознавателя, относится сразу к нескольким классам визем. Это позволяет контролировать и использовать промежуточные, переходные состояния между виземами, что может быть очень полезно при решении задачи автоматического анализа последовательности визем – слов или предложений в процессе слитной речи.

Каждое изображение базы данных визем было дополнено зеркально отражённым в горизонтальном направлении изображением. Для уменьшения влияния масштаба изображения на входе на значения откликов

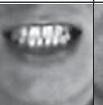
прямоугольных свойств (см. рис. 5) изображения в процессе обучения дополнялись набором тех же изображений различного масштаба. Для экспериментов использовались 30 масштабов обучающих изображений в диапазоне 30–90 пикселей с шагом 2 пикселя. Используемый диапазон масштабов является характерным для входных изображений области губ, получаемых на выходе алгоритма поиска губ. Для объективного тестирования полученных результатов обучения используемая база изображений визем была разделена на две части – обучающий и тестовый наборы. На обучающем наборе проводилось обучение классификаторов. На тестовом наборе, не пересекающемся с обучающим, проводилось тестирование полученных классификаторов.

Процесс обучения по методу AdaBoost проводился до достижения классификатором средней ошибки классификации менее 0.001. Количество используемых прямоугольных признаков, необходимых для достижения указанной ошибки, в экспериментах не превосходило 50. С учётом того, что входными данными для распознавания является область изображения, заключающая в себе предварительно обнаруженные губы, можно сделать вывод о высокой скорости обработки данных и об отсутствии необходимости её увеличения за счёт использования каскада классификаторов.

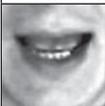
Тестирование полученных классификаторов на тестовом наборе показало результаты, приведённые в табл. 4.

Таблица 4

Матрица принятых решений на тестовом наборе

Входные виземы	Распознанные виземы									Множественное распознавание или отказ от распознавания
	1	2	3	4	5	6	7	8	9	
	a	o	y	e	i	п	ф	ш	\$	
					и	б	в	ж		
						м		ч		
								дж		
										
	66	-	-	2	-	-	-	-	-	12
	-	76	-	-	-	-	-	-	-	8
	-	-	103	-	-	-	-	-	1	8



Входные виземы	Распознанные виземы									Множественное распознавание или отказ от распознавания
	1	2	3	4	5	6	7	8	9	
	а	о	у	е	і	п	ф	ш	џ	
	-	-	-	28	1	-	-	2	-	3
	-	-	-	-	14	-	-	-	-	-
	-	-	-	-	-	38	-	-	2	6
	-	-	1	-	-	-	41	-	-	2
	-	-	-	-	-	-	-	33	-	9
	-	-	-	-	-	-	-	-	93	7

Приведённые в табл. 4 данные говорят о достижении удовлетворительных результатов по распознаванию визем. Наибольшие ошибки связаны с множественным распознаванием и отказом от распознавания, что объясняется влиянием индивидуальных визуальных особенностей артикуляционного аппарата различных людей. В приведённой матрице принятых решений, полученной на тестовом наборе, не выявлено устойчивой неправильной классификации каких-либо двух классов, что свидетельствует об отсутствии необходимости сокращения принятого алфавита классов-визем путём объединения плохо разделяемых классов в один. Достигнутые показатели качества позволили использовать описанный подход при разработке программы обучения правильной артикуляции для облегчения восприятия устной речи людьми с нарушениями слуха.

Литература

1. Нейман Л.В. Анатомия, физиология и патология органов слуха и речи. М.: Просвещение, 1977 г. 146 с.
2. Методика обучения глухих устной речи: Учеб. пособие для студентов дефектол. фак. фед. ин-тов. / Под ред. проф. Ф.Ф. Рау. М.: Просвещение, 1976. 279 с.
3. <http://www.pedlib.hut.ru/Books/pravdina/pravdinap124.html>.
4. Мурыгин К.В. Концепция системы распознавания речи на основе чтения по губам // Искусственный интеллект. 2009. №2. С. 116–123.

5. Paul Viola and Michael J. Jones. Robust real-time object detection //In Proc. of IEEE Workshop on Statistical and Computational Theories of Vision, 2001.
6. Мурыгин К.В. Обнаружение объектов на изображении на основе каскада классификаторов // Искусственный интеллект. 2007. №2. С. 104–108.
7. Мурыгин К.В. Особенности реализации алгоритма AdaBoost для обнаружения объектов на изображениях // Искусственный интеллект. 2009. №3. С. 573–581.

Мурыгин К.В.

*Кандидат технических наук, начальник отдела распознавания зрительных образов
Института проблем искусственного интеллекта (г. Донецк, Украина).*

*Область научных интересов: разработка методов и алгоритмов интеллектуального
анализа зрительной информации. Занимался вопросами поиска объектов на
изображении, распознавания человека по изображению лица, анализа движения
в видеопоследовательности, стереозрения – восстановления карты диспаратности
по набору изображений.*

info@iai.dn.ua



Озвучивание SMS-сообщений, отправляемых на стационарные телефоны

*Людовик Т.В.,
кандидат технических наук
Робейко В.В.*

Тексты SMS-сообщений отличаются спонтанностью, экспрессивностью, открытым характером лексикона, а также зачастую многоязычностью и неправильной орфографией. В данной статье описывается опыт использования системы синтеза украинской речи для озвучивания SMS-сообщений, посылаемых на стационарные телефоны. Описывается подход, позволяющий синтезировать как нейтральную, так и выразительную речь различной коммуникативной ориентации. Результаты проведённого тестирования свидетельствуют о том, что потенциальные пользователи предпочитают, чтобы SMS озвучивались выразительным женским голосом.

SMS texts are characterized by spontaneity, expressivity, open lexicons, multilinguality, and incorrect spelling. In this article we describe the experience of applying a TTS system for Ukrainian language to reading SMS texts sent to fixed phones. We describe the approach allowing to synthesize both neutral and expressive speech with distinct communicative orientation. Test results show that people prefer SMS texts synthesized with an expressive female voice.

Введение

В данной статье излагается опыт, приобретённый в рамках внедрения системы синтеза украинской речи [1] в службу, обеспечивающую автоматическое озвучивание SMS-сообщений.

Сервис SMS2Voice (SMS2Fix), разработанный немецкой компанией, которая занимается информационными коммуникационными технологиями [2],

позволяет пользователям сетей различных мобильных операторов посылать SMS обычного формата на стационарные телефоны. Данный сервис внедрён в Украине украинской компанией Global Message Services Ukraine (GMSU). На данный момент обслуживаются три языка: украинский, русский и английский. Соответственно, задействованы три различных синтезатора речи. Для русского и английского языков используются разработки известных иностранных компаний. Для украинского языка используется синтезатор речи, созданный украинской компанией CyberMova с учётом разработок Международного научно-учебного центра информационных технологий и систем (г. Киев).

Статья посвящена приложению системы синтеза украинской речи к озвучиванию специфических текстов (SMS) в специфических условиях (однократное прослушивание абонентом сообщения, озвученного незнакомым голосом). Мы не будем рассматривать такие важные и сложные проблемы, как автоматическое распознавание языков и транслитерация текстов SMS латиницей, поскольку первая из этих задач решается немецкими коллегами, а вторая – совместно с ними. Также мы не будем подробно останавливаться на решениях, связанных с синтезом речи на русском языке. Однако будут затронуты проблемы, общие для украинского и русского языков в связи с их тесным переплетением в мобильном пространстве Украины.

Типичный пример озвученного SMS-сообщения выглядит следующим образом:

Добрий вечір. (*Добрый вечер.*)
Ви отримали SMS-повідомлення. (*Вы получили SMS-сообщение.*)
Я прийду до дому з нивестой.
Відправник (*Отправитель*): +38 097 9897931.
Дата відправлення (*Дата отправки*): 12.10.2009. 11:00.
На все добре. (*Всего доброго.*)

По приведённому примеру можно судить о том, насколько проблема озвучивания текстов усложняется наличием орфографических и иных ошибок: правильный текст SMS должен выглядеть так: «Я прийду додому з нареченою» («Я приду домой с невестой»).

Исследованный материал

Материал в виде реальных SMS, отправленных в период с 17 мая 2008 года по 6 июня 2008 года, был анонимизирован путём исключения данных об отправителях и получателях. Таким образом, анализировались лишь тексты сообщений. Экспертом было выделено пять категорий SMS-текстов в соответствии с языком, на котором они написаны. В таблице 1 представлено распределение проанализированных SMS-сообщений.

В дальнейшем проводился анализ первой и второй категорий. Следует отметить, что так называемый украинско-русский язык не является ни украинским языком с «русскими» включениями, ни наоборот. Анализ реальных SMS свидетельствует о том, что в Украине распространён «суржик» – гибрид украинского и русского языков, усложняющий синтез речи на обоих языках и требующий изучения специалистами.



Таблица 1

Распределение текстов SMS-сообщений

Язык	Количество SMS (%)	Примеры
Украинский	1523 (14.1%)	Як справи? (Как дела?)
Украинско-русский («суржик»)	1262 (11.7%)	Умееш находить похидну с корня? (производную) Мама, готов шось поесть. Напиши SMS.
Русский	7640 (70.6%)	Я на работе.
Другие языки	274 (2.5%)	Tutto ok. E mille grazie
Нетекстовые SMS	126 (1.2%)	@1@2@3@

Проблемы синтеза речи, связанные со спецификой SMS-текстов

Наряду с проблемами повышения разборчивости (актуальной в условиях использования телефонных каналов связи) и естественности синтезированной речи, адаптация системы синтеза украинской речи к озвучиванию SMS потребовала решения дополнительных задач из области социо- и психолингвистики.

В таблице 2 перечислены основные проблемы, связанные с озвучиванием SMS-сообщений.

Как видно из примеров, многоязычность SMS-текстов может проявляться не только на уровне слов, но и на уровне букв, что существенно затрудняет автоматическое распознавание языков и расстановку словесного ударения, из-за того что неправильно написанные слова отсутствуют как в украинском, так и в русском словарях.

Как правило, SMS-тексты имеют выраженную коммуникативную направленность, которую желательно отразить при их озвучивании. Основным средством для этого является интонация. В настоящее время ставится задача моделирования декларативного, восклицательного и вопросительного коммуникативных типов в синтезированной речи. В частности, общий и специальный вопросы должны отличаться интонационно.

В отличие от проблемы выражения в озвучиваемых SMS эмоциональности и спонтанности, решение которой представляется делом будущего, проблемы, связанные с омографией, расшифровкой аббревиатур и сокращений, а также с отсутствием в текстах SMS разделительных знаков, должны решаться уже сегодня.

Таблица 2

Основные проблемы, связанные с озвучиванием SMS-сообщений

Специфические проблемы озвучивания SMS-сообщений	Примеры
Многоязычность	Вы хто? Сонечко, мы с тобой встречаемся? Прівэт бабуля Проверка св'язі ! Шо Робыш?
Неправильная орфография: а) намеренная; б) описки; в) неграмотность	Яаа люблю тильки тебее! Дякою. Візьми тіліфон. Яанемагу.позванити.поытому.номиру.
Коммуникативная направленность	Бабуся я тебе люблю! Вітаю зі святом! Батьки збираються? До тебе можна? Доброго дня!Як настрої?А можна Інну до телефону? Нехай росте вам на радість! Хай!Як справи?Це Таня з Деревич це мій ном.пиши що там Руся,Ярик? Галя, нельзя так себя вести. Давай вставай. Давай когда?
Эмоциональность	это я!!! Ааааа!!!! Я заклеил гитару!! Поставил струны и она просто ВЕЛИКОЛЕПНО звучит! Алла!!! Шабаш Давай!!! Дима ты не послушный!!! Ж Д У !!! Хи-хи хи-хи.как делифки??? Игорь вставай!!! Лиля! Делай уроки! Немедленно!ясно???! Я їду з Дімою!!! де мое бабло вчорашне??? ЗРОЗУМІТИ ТЕБЕ! А ТИ ЗРОЗУМІЛА МЕНЕ??? Ты ведь сам хотел розстатса??? Я права???! АЛЕ ЦЕ ВЖЕ ЗАНАДТО! НЕ БРАТИ ТРУБКУ КОЛИ Я ДЗВОНЮ! ТИ ПЕРЕГИНАЄШ ПАЛКУ!
Отсутствие знаков пунктуации	Як дела що в тебе нового у мене всьо постарому привет вид дивчат. Лераятебялюблюіочуштобтиприехалазамной АтакВсеДобре. ГОВОРИ ГОВОРИ А ТЕБЕ КАК ОБ СТЕНКУ ГОЛОВОЙ ДУМАЙ И ВСПОМНИ
Омография и омофония	Бажая добра і тепла. АНЮТКА,МЫ НЕКУДА НЕ ЕДИМ.
Аббревиатуры и сокращения	візьми 2 грн на завтра Вол. Ів. Дякую. Виба4. г.Одессы Что сег.собир.делать?



Подход к созданию женского выразительного голоса

SMS на всех трёх языках озвучиваются методом Unit Selection [3], подразумевающим использование больших баз данных, содержащих отрезки реальной речи и отражающих как особенности голосов конкретных дикторов, так и стиль чтения, используемый диктором. В первом коммерческом варианте системы синтеза украинской речи использовался мужской голос ПАНАС с довольно монотонным стилем чтения художественной литературы. Считается, что для озвучивания SMS больше подходит женский голос, при этом синтезированная речь должна звучать в нейтральном стиле или в стиле «хорошие новости» [4–7].

Новый женский голос НАТАЛКА, используемый в настоящее время в рамках сервиса «SMS на стационарный телефон», отличается от мужского большей естественностью и выразительностью.

Как и ранее, для синтеза украинской речи применяется метод Unit Selection [3] в комбинации с индивидуализированными просодическими моделями.

Применяемый подход позволяет синтезировать естественно звучащую речь с индивидуальными интонацией и ритмикой.

Речевая база данных

Речевая база данных является одним из основных компонентов системы синтеза речи методом Unit Selection. Как правило, используется голос профессионального диктора, записанный в студийных условиях. В зависимости от сферы будущего применения синтезированной речи диктор читает различные подготовленные тексты (новости, прогнозы погоды и т.п.).

При создании голоса НАТАЛКА, предназначенного, в частности, для озвучивания SMS-сообщений, были отобраны и составлены тексты различных стилей. Чтение этих текстов профессиональным диктором-женщиной должно было отразить как различную коммуникативную направленность текстов (чтение художественной литературы, новостей, диалогов, SMS-сообщений), так и общепринятое произнесение телефонных номеров, дат, электронных адресов и т.п.

В таблице 3 описан акустический материал, послуживший основой речевой базы данных голоса НАТАЛКА.

Объединение в одной речевой базе данных акустических записей, неизбежно отличающихся степенью выразительности при использовании текстов разных типов, порождает дополнительную сложность при синтезе речи. С одной стороны, это позволяет синтезировать более выразительную речь; с другой стороны, затрудняется процесс поиска речевых элементов в базе данных.

Использованные тексты были затранскрибированы автоматически. Поскольку произношение диктора нормативное, результаты автоматического фонем-

Таблица 3

**Речевой материал, послуживший основой
для создания голоса НАТАЛКА**

Тип текста	Количество слов в тексте	Продолжительность записи для данного текста (мин.)	Продолжительность записи для данного текста (%)
Художественный текст «Алиса в Стране Чудес»	10602	130	65
Тексты SMS-сообщений	619	8	4
Диалоги	208	2	1
Новости	478	5	2
Украинский и английский алфавиты	61	2	1
Названия сайтов, даты и т.п.	280	5	2
Специально сконструированные фразы	3483	49	24
Всего	15731	201	100

ного транскрибирования не потребовали коррекции. Вручную в транскрипцию были вставлены паузы, добавленные диктором и не соответствующие знакам пунктуации. Полученная транскрипция была использована для автоматической фонемной сегментации речевого материала с помощью пакета программ НТК [8]. Границы аллофонов были откорректированы с помощью автоматизированных средств обработки речевых сигналов [9]. Дальнейшая сегментация выделенных аллофонов на периоды основного тона была произведена в автоматическом режиме с незначительной ручной коррекцией.

В речевой базе данных, насчитывающей около 80 000 аллофонов, каждый элемент аннотирован: идентификатором, именем, состоящим из трёх частей (имя предыдущей, текущей и последующей фонем), длительностью всего аллофона, уровнями интенсивности первой и второй половин, а также для гласных и звонких согласных – последовательностью длительностей периодов основного тона и количеством периодов. Таким образом, в аннотации речевой базы данных приведена исключительно объективная информация, т.е. отсутствуют символные просодические метки [10].

Предобработка текстов SMS-сообщений

Задача озвучивания SMS-сообщений потребовала существенного развития модуля предобработки текстов [2]. В настоящее время осуществляются такие преобразования:

- описание эмотиконов типа :) соответствующими словами («смайлик»);
- замена слов из списка «неприличная лексика» звуковым сигналом «би-и-и-п»;
- транслитерация SMS-сообщений, написанных латиницей;
- обработка обозначения дат, телефонных номеров, адресов web-страниц и т.п.;
- членение длинных предложений на синтагмы;
- расстановка словесных ударений;



- расшифровка аббревиатур и сокращений;
- преобразование чисел и символов в орфографический текст.

Особое внимание было уделено расстановке ударений и расшифровке аббревиатур. Словари, используемые для этих операций, имеют открытый характер и динамически пополняются (обновляются). Тем не менее, вероятность ненахождения некоторых слов в словарях остаётся высокой из-за креативности или неграмотности отправителей SMS.

Расстановка ударений связана также со снятием омографии. Эта проблема продолжает оставаться сложной, поскольку система синтеза речи не производит семантический и прагматический анализ текстов.

Транскрибирование текстов SMS с учётом особенностей произношения диктора

Модуль транскрибирования был расширен введением правила вставки гортанной смычки перед фонетическими словами, начинающимися гласной фонемой. Гортанная смычка в этой позиции характерна для голоса НАТАЛКА. Кроме того, выполняя функцию граничного маркера [11], гортанная смычка способствует повышению качества синтезированной речи, улучшая её разборчивость.

Вычисление просодических характеристик SMS-сообщений

Вычисление длительности синтезируемых звуков осуществляется с помощью простой модели, параметрами которой являются средняя длительность фонемы, тип контекста, в котором она находится в синтезируемом высказывании, и набор коэффициентов длительности для данной фонемы. Модель длительности звуков индивидуализируется оффлайн на основе автоматического анализа аннотации речевой базы данных.

Разработанная модель интонирования близка к модели интонационных портретов акцентных единиц, предложенной Б.М. Лобановым [12].

Модель интонирования также была индивидуализирована с учётом особенностей голоса НАТАЛКА. На рисунке 1 приведены стилизованные интонационные контуры завершенности и незавершенности, используемые при синтезе речи голосами ПАНАС и НАТАЛКА. Интонационные фразы (синтагмы), представленные на рисунке, состоят из трёх акцентных групп (АГ). Представлен наиболее типичный случай, когда последняя акцентная группа синтагмы несёт синтагматическое ударение. Приведённые контуры нормированы. Ось абсцисс соответствует времени, а ось ординат – нормированным значениям частоты основного тона. Нормированное значение частоты основного тона «0» соответствует 150 Гц для голоса НАТАЛКА и 80 Гц для голоса ПАНАС. Максимальное значение «10» соответствует 375 Гц для голоса НАТАЛКА и 180 Гц для голоса ПАНАС. На рисунке видно, что изменения интонационных контуров голоса НАТАЛКА более резко выражены. Это свидетельствует о большей экспрессивности этого голоса.

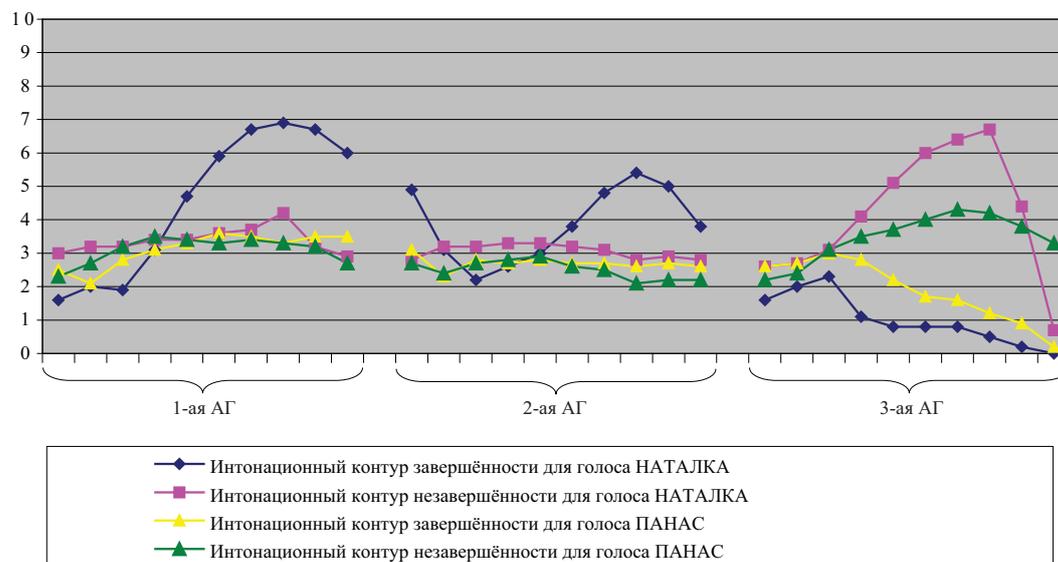


Рис. 1. Стилизованные интонационные контуры завершенности и незавершенности для женского голоса НАТАЛКА и для мужского голоса ПАНАС

Поиск в базе данных

Описанные выше модули вырабатывают целевое описание озвучиваемого текста – его фонемно-просодическую транскрипцию, в которой указана последовательность фонем, их длительности и их интонационные контуры в виде значений длительности периодов основного тона.

Алгоритм выбора аллофонов (Unit Selection) из речевой базы данных НАТАЛКА отличается от предыдущего варианта, предназначенного для голоса ПАНАС. В частности, используются дополнительные критерии выбора, связанные с интенсивностью (громкостью) аллофонов. Первоначальный алгоритм выбора работал локально, т.е. учитывалась не вся цепочка аллофонов синтагмы, а только текущий выбираемый аллофон и аллофон, выбранный на предыдущем шаге. Для голоса НАТАЛКА был разработан новый вариант алгоритма Unit Selection, приближенный к классической схеме [3].

Генерация акустического сигнала

Генерация акустического сигнала представляет собой конкатенацию речевых отрезков, соответствующих аллофонам, выбранным из речевой базы данных. При синтезе речи голосом ПАНАС конкатенация осуществляется с просодической модификацией ударных гласных (изменением их длительности и контура основного тона). Аллофоны голоса НАТАЛКА не модифицируются.

Результаты

Было проведено формальное тестирование с целью определить, какой из голосов (ПАНАС или НАТАЛКА) звучит качественнее (естественнее). В частности, обоими голоса-



ми было озвучено 33 SMS-текста, т.е. было синтезировано 66 стимулов. Эксперимент проходил в три этапа. На первом этапе аудиторами были 5 человек – сотрудников МНУЦИТиС, на втором – 33 преподавателя и студента-лингвиста (специальность «украинский язык и литература»), на третьем – 4 сотрудника компании Global Message Services Ukraine. Всем аудиторам предлагалось прослушать синтезированные стимулы, предъявляемые в случайном порядке, и выставить каждому стимулу оценку от 0 (очень плохо) до 5 (очень хорошо).

Результаты тестирования приведены на диаграмме (рис. 2). Все три группы аудиторов отдали предпочтение женскому голосу НАТАЛКА.

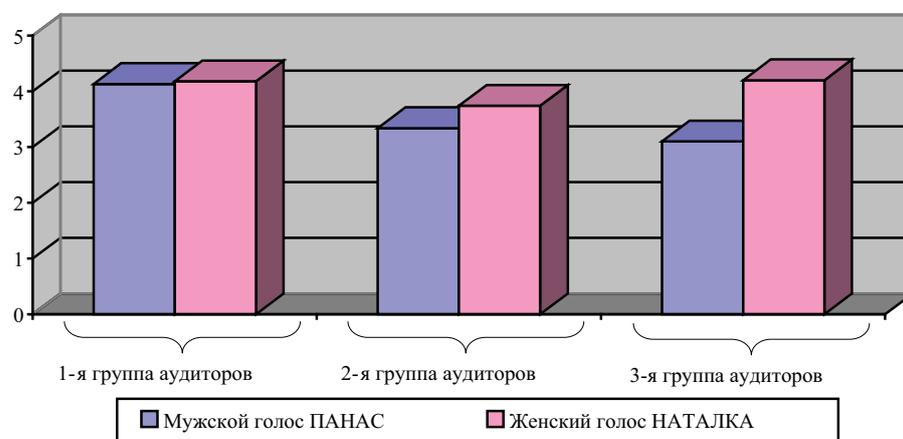


Рис. 2. Результаты тестирования естественности озвученных SMS

Заключение

Опыт применения системы синтеза речи к озвучиванию SMS-сообщений свидетельствует о том, что традиционная для синтеза речи проблема повышения естественности звучания продолжает оставаться актуальной. При этом сложно найти компромисс между нейтральным и выразительным (эмоциональным) стилями озвучивания. Очевидны также социолингвистические аспекты проблемы озвучивания SMS-сообщений, связанные с многоязычностью и широким распространением «суржика». Отдельной проблемой является открытый характер словарей, требующий постоянного мониторинга.

Литература

1. Lyudovyyk T., Sazhok M. Unit Selection Speech Synthesis Using Phonetic-Prosodic Description of Speech Databases, Proc. International Conference «Speech and Computer» (SPECOM'2004), St.-Petersburg (Russia), 2004. P. 594–599.
2. Lyudovyyk T., Brozinski S., Noner M., Robeiko V., Sazhok M. Speech Synthesis Applied to SMS reading, Proc. of the 13th International Conference «Speech and Computer: SPECOM'2009». – St.-Petersburg, Russia, 2009. P. 300–305.

3. *Hunt A., Black A.* «Unit selection in a concatenative speech synthesis system using large database», Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1996. P. 373–376.
4. *Shinsuke S., Jinfu N., Ranniery M., Keiichi T., Minoru T., Tomoki Toda, Hisashi Kawai and Satoshi Nakamura.* Communicative Speech Synthesis with XIMERA: a first step, Proc. of 6th ISCA Speech Synthesis Workshop, Bonn, Germany, August 22–24, 2007 (CD-ROM proceedings).
5. *Fernandez R., Ramabhadran B.* Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis. In Proceedings of the 6th ISCA Workshop on Speech Synthesis, pages 34–39, Bonn, Germany, 2007.
6. *E. Eide, A. Aaron, R. Bakis, W. Hanza, M. Picheny and J. Pirelli.* A corpus-based approach to <AHEM/> expressive speech synthesis, in Proceedings of 5th ISSW, June 2004. P. 79–84, Pittsburgh, USA.
7. *Theune M., Meijs K., Heylen D.K.J., Ordelman R.J.F.* Generating Expressive Speech for Storytelling Applications, IEEE transactions on audio, speech and language processing, 2006, 14 (4). P. 1137–1144.
8. *Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Valtchev V., Woodland P.* The HTK Book, Cambridge University Engineering Department, 2002.
9. *Сажок М.М.* Автоматизовані засоби формування баз даних і знань для озвучення українських текстів: Автореф. дис. канд. техн. наук: 05.13.06. К.: МННЦІТІС, 2004. 20 с.
10. *Silverman K., Beckman M., Pitrelli J.* ToBI: a Standard for Labeling English Prosody, Proc. of the International Conf. on Spoken Language Processing. Banff, 1992. Vol.2. P. 867–870.
11. *Кривнова О.Ф.* Ларингализация как граничный маркер в устной речи, Труды XVI сессии Российского акустического общества РАО-16. М. 2005.
12. *Лобанов Б.М., Цирульник Л.И.* Компьютерный синтез и клонирование речи. Минск: Белорус. Наука, 2008. 337 с.

Людвик Татьяна Владленовна

Кандидат технических наук.

Работает старшим научным сотрудником Международного научно-учебного центра информационных технологий и систем (г. Киев, Украина). Закончила Киевский государственный университет по специальности «Прикладная лингвистика». Главные научные интересы лежат в области синтеза речи и экспериментальной фонетики.

Робейко Валентина Васильевна

Закончила Институт филологии Киевского национального университета им. Т. Шевченко. Специалист в области современного украинского языка. Работает младшим научным сотрудником в Международном научно-учебном центре информационных технологий и систем (г. Киев, Украина), аспирантка. Область интересов: синтез и распознавание речи, фонетика.



Опыт автоматического стенографирования украинской парламентской речи



Пилипенко В.В.

Робейко В.В.

В статье рассматривается программа-стенограф для получения текста стенограммы из звукового файла на основе системы распознавания речи. Записанная фонограмма обрабатывается системой распознавания слитной речи многих дикторов с большими словарями (больше 10 тыс. слов). Оператор исправляет допущенные ошибки, а также вводит новые слова, незнакомые системе распознавания. На основе анализа ошибок и новых слов производится дообучение системы распознавания, что позволяет улучшать показатели надёжности распознавания речи в процессе эксплуатации системы. Применение индивидуализированных транскрипций для некоторых дикторов позволило повысить надёжность распознавания. Приведены результаты распознавания украинской парламентской речи.

In this work we present a research system of computerized stenographer. It produces texts out of sound records by means of human-aided speech recognition system. Continuous speech recognition system for a number of speakers with a large vocabulary (more than 10.000 words) is used to process recorded files. Human introduces out-of-vocabulary words and repairs errors to produce the final text. Retraining process is running to take into account repairs, thus improving system performance. Personal phonetic rules are listed and used to individualize transcriptions for different speakers. Experimental results for Ukrainian parliament speech recognition are presented.

Список ключевых слов

Распознавание слитной речи, автоматизированное стенографирование, многодикторное распознавание речи, большие словари, индивидуализированные транскрипции.

Continuous speech recognition, computer-assisted stenographer, multispeaker speech recognition, large vocabulary, transcription individualization.

Введение

Стенографирование широко используется для обработки и документирования материалов заседаний и совещаний различного уровня, для работы секретарей, журналистов и так далее. Во многих странах необходимо стенографировать заседания в парламентах и судах. Компьютеры значительно расширили возможности и увеличили гибкость применения систем стенографирования. На данный момент становится актуальной задача уменьшения доли ручного труда в таких системах. Для этого предлагается использовать автоматическое распознавание речи при преобразовании звука в текст.

Речь каждого человека сугубо индивидуальна. Поэтому перевести звук в текст по нажатию одной кнопки — задача довольно сложная для системы стенографирования. Такая система должна максимально упростить работу оператора и ускорить перевод звукового файла в текстовый, а также учесть все особенности речи диктора. Существует много программно-аппаратных комплексов автоматизированного стенографирования с различными возможностями, но даже самый простой позволяет увеличить скорость перевода звука в текст в несколько раз.

Автоматическое распознавание слитной речи многих дикторов с большими словарями значительно упрощает работу оператора, сводя её к исправлению ошибок, допущенных системой стенографирования. Дообучение системы позволяет сокращать количество ошибок в процессе эксплуатации.

1. Автоматизированная vs автоматическая

Системы стенографирования можно условно разделить на три категории в зависимости от соотношения участия человека и компьютера в процессе создания стенограмм:

- **автоматические** (без участия человека в процессе распознавания речи);
- **автоматизированные** (компьютер распознаёт поток речи, а человек участвует в этом процессе в той или иной степени);
- **стенографирование при помощи компьютера** (человек набирает текст, а компьютер используется как магнитофон и печатная машинка).

Разница между автоматической и автоматизированной системами заключается в надёжности автоматического распознавания речи.

Опыт эксплуатации показывает, что первичная стенограмма, созданная человеком, содержит ошибки, которые исправляются в процессе редактирования набранного текста. В среднем количество ошибок достигает 5 на одну страницу текста, что составляет надёжность 98%, поскольку одна страница содержит приблизительно 2000 знаков, или 250 слов. Таким образом, система стенографирования должна стать автоматической при надёжности распознавания речи выше 98%.

Такая надёжность уже сегодня достижима для автоматического распознавания речи при некоторых ограничениях [1]. При этом распознаётся речь только одного диктора. Для



изолированно произносимых слов словарь достигает 15 тыс. слов, а для слитной речи такая надёжность достигается при словаре в 1 тыс. слов.

Поэтому на настоящий момент актуальным является создание программ распознавания речи, свободных от таких ограничений. Для стенографирования необходимо достигнуть объёмов словаря от 10 тыс. слов до нескольких миллионов. Количество задействованных дикторов составляет от 100 до нескольких тысяч. При этом должна распознаваться слитная речь в реальном времени.

Автоматизированную систему стенографирования имеет смысл применять при надёжности 80% и выше. При этом оператору необходимо будет исправлять не более чем каждое пятое слово в тексте, что можно делать при прослушивании звуковой дорожки в процессе её воспроизведения.

2. Система распознавания слитной речи

В данной работе в качестве базовой системы используется инструментарий НТК [2] на основе скрытых Марковских моделей (СММ). Инструментарий НТК использовался для построения акустических и лингвистических моделей. Для распознавания речи был разработан программный комплекс, совместимый с акустическими и лингвистическими моделями НТК.

2.1. Пользовательский вид программы

Пользовательский вид программы стенографирования приведён на рисунке 1. В верхнем окне схематически изображена осциллограмма звуковой

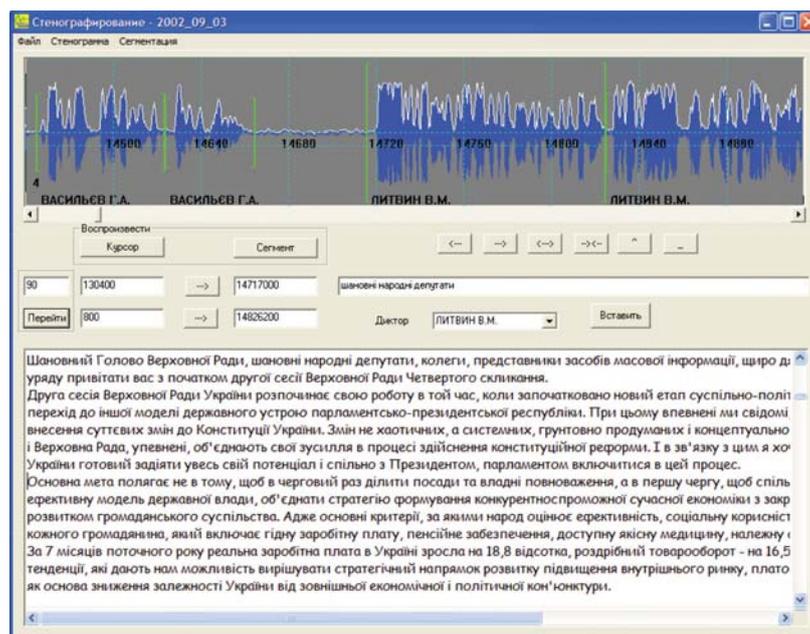


Рис. 1. Общий вид программы стенографирования

дорожки с автоматически выделенными сегментами речи (фразами или синтагмами). Оператор выделяет нужный ему сегмент и прослушивает его. При этом он может посмотреть ответ распознавания, который можно исправить в случае ошибки. После редактирования ответ добавляется в стенограмму и автоматически происходит переход к следующему сегменту.

Пользователь имеет возможность перейти к нужному диктору или прослушать необходимый сегмент стенограммы.

Распознавание производится автоматически в фоновом режиме работы программы. Все ошибки распознавания фиксируются и, после того как закончилось формирование стенограммы, используются для дообучения системы стенографирования. При этом в обучающую выборку добавляются новые слова и информация о новых дикторах. Таким образом, надёжность распознавания повышается в процессе эксплуатации системы стенографирования.

2.2. Предварительная обработка речевого сигнала

Речевой сигнал, оцифрованный на частоте 22050 Гц с точностью 16 бит, преобразуется в последовательность векторов признаков с интервалом анализа 25 мсек. и шагом анализа 10 мсек. Вначале речевой сигнал фильтруется фильтром высоких частот с характеристикой $P(z) = 1 - 0.97z^{-1}$. Затем применяется окно Хэмминга и вычисляется быстрое преобразование Фурье. Спектральные коэффициенты усредняются с использованием 26 треугольных окон, расположенных в мел-шкале, и вычисляются 12 кепстральных коэффициентов.

Логарифм энергии добавляется в качестве 13-го коэффициента. Эти 13 коэффициентов расширяются до 39-мерного вектора параметров путём дописывания первой и второй разностей от коэффициентов, соседних по времени. Для учёта влияния канала применяется вычитание среднего кепстра.

2.3. Акустическая модель

В качестве акустических моделей используются скрытые Марковские модели. 56 украинских фонем (включая фонему-паузу) моделируются тремя состояниями Марковской цепи без пропусков. Используется диагональный вид Гауссовских функций плотности вероятности.

Редко встречающиеся фонемы моделируются 64 смесями Гауссовских функций плотности вероятности, более часто встречающиеся фонемы моделируются большим числом смесей, наиболее часто встречающиеся фонемы используют 1024 смеси.

Словарь транскрипций создаётся автоматически из орфографического словаря с использованием контекстно-зависимых правил.

2.4. Многодикторная система

Распознавание речи независимо от диктора является задачей, не решённой до конца в распознавании речи. В [3] использовалась модель кооперативного распознавания, в



которой при обучении смешивалась речь разных дикторов. При этом речь разных дикторов рассматривалась как разные реализации речи одного диктора. Было показано, что надёжность распознавания улучшалась не только для дикторов, входящих в кооператив, но также и для дикторов, незнакомых системе. Скорее всего, это связано с тем, что речь знакомых системе дикторов похожа на речь других дикторов.

Опыт применения такого подхода показал, что при использовании больше 100 дикторов в кооперативе надёжность распознавания речи становится очень близкой к системе, независимой от диктора.

Методы работы с многими дикторами, заложенные в инструментарий НТК, такие, как нормализация длины речевого тракта (Vocal Tract Length Normalisation) и адаптация модели при помощи линейного преобразования максимального правдоподобия (Maximum Likelihood Linear Regression), позволяют улучшить надёжность распознавания речи для отдельных дикторов при условии, что каким-то независимым способом заранее определяется диктор. Предполагается использовать методы идентификации дикторов для автоматического определения говорящего.

3. Акустическое и текстовое наполнение

3.1. Обучающая выборка

Обучение производилось на выступлениях депутатов Верховной Рады Украины, записанных через телевизионную сеть. Парламентская речь характеризуется некоторыми особенностями.

- Это спонтанная речь. Встречаются отдельные доклады, зачитываемые по подготовленному заранее тексту, однако мало дикторов в точности придерживается этого текста.
- Из-за ограничения во времени выступления многих дикторов произносятся в слишком быстром темпе.
- Часто речь эмоционально окрашена.
- В основном, записи состоят из непрерывных выступлений дикторов, но в них встречаются реплики ведущего заседания или других депутатов.
- Качество записи достаточно высокое, поскольку каждое депутатское место оснащено микрофоном.

Для обучения использовались записи длиной в 197 тыс. секунд, в которых встретилось 427 576 слов. Всего было записано 287 дикторов. Дикторов с длиной больше 300 сек. оказалось 139, что позволяет сформировать качественные акустические модели для данных дикторов. Время записи сильно отличается для разных дикторов.

Обучение производилось на предварительно размеченной выборке. Для этого запись выступления автоматически разбивалась на фразы из нескольких слов, ограниченные паузами больше 400 мсек. Каждой фразе оператор ставил в соответствие метку в виде текста из стенограммы. Затем автоматически производилось преобразование текста в последовательность фонем в соответствии с контекстно-зависимыми правилами. Выборка, размеченная таким образом, использовалась для построения акустической модели при помощи инструментария НТК.

3.2. Контрольная выборка

Распознавание производилось на выступлениях депутатов, записанных в дни, отличающиеся от обучающей выборки. Для распознавания использовались записи длиной в 42 тыс. секунд, в которых встретилось 94 817 слов. Всего использовались записи 152 дикторов. Дикторов с длиной больше 300 сек. оказалось 45. Записи 41 диктора не встретились в обучающей выборке. Таким образом, эти дикторы оказались незнакомыми для системы распознавания.

3.3. Текстовый материал

Словарь был составлен из текстов стенограмм заседаний Верховной Рады Украины. С официального сайта Верховной Рады были загружены все стенограммы заседаний, начиная с 1991 года, что составило больше 100 МБ текста. Текст был модифицирован для того, чтобы убрать служебную информацию из стенограмм (например, обозначения аплодисментов и т.п.), записать числа в текстовом виде, а также отделить русский текст от украинского.

Результирующий текст разделён на две части: первая (обучающая выборка) содержит все тексты, кроме 2002—2003 годов, вторая содержит стенограммы 2002—2003 годов (контрольная выборка). Первая часть состоит из 14 629 111 словоупотреблений, во второй содержится 409 244.

Для первой части был составлен словарь из 156 108 слов и подсчитаны частоты встречаемости слов в этом словаре. Таблица 1 показывает долю текстов, которая покрывается словарём с определённой частотой встречаемости слов. Видно, что весь словарь (156 108 слов, частота встречаемости слов — один раз и больше) покрывает 99,6% нового текста. Доля текста из контрольной выборки, покрываемого словами с частотой встречаемости в обучающей выборке 50 раз и больше, составляет 94,9%. Для этого достаточно иметь словарь объёмом в 15 тыс. слов.

Исследовалась надёжность распознавания в зависимости от объёма частотного словаря с использованием биграммной модели языка. Результаты представлены в таблице 2, из которой видно, что надёжность незначительно увеличивается при увеличении размера словаря. «Закрытого» словаря объёмом в 15 тыс. слов достаточно для распознавания речи с небольшим (2%) уменьшением надёжности по сравнению с максимально возможным словарём. «Закрытый» словарь не предполагает возможность отказа от распознавания.



Таблица 1

Доля текстов, входящих в частотные словари

Частота встречаемости слов	Количество слов в частотном словаре	Встречаемость в текстах обучающей выборки	%	Встречаемость в текстах контрольной выборки	%
1	156 108	14 629 111	100,0	407 608	99,6
2	98 601	14 571 604	99,6	406 563	99,3
3	78 022	14 530 446	99,3	405 753	99,1
5	58 936	14 465 646	98,8	404 293	98,7
10	40 364	14 343 499	98,0	401 544	98,1
50	15 609	13 805 357	94,3	388 470	94,9
100	10 032	13 415 092	91,7	378 873	92,5
200	6 219	12 878 264	88,0	365 510	89,3
300	4 622	12 488 459	85,3	355 606	86,8

Таблица 2

Надёжность распознавания для разных объёмов частотного словаря

Объём словаря	64 000	50 000	30 000	20 000	15 000	10 000	5 000
Надёжность распознавания (%)	68,59	68,54	68,38	67,79	67,15	65,49	62,18

4. Биграммная модель языка

При распознавании речи использовалась биграммная модель речи, которая задавалась вероятностями появления пар слов. Поскольку в текстах, на которых вычислялись статистики, встретились далеко не все пары слов, возможные для данного словаря, то для аппроксимации *ненаблюдаемых* пар слов использовались так называемые *обратные* (back off) коэффициенты [2].

Биграммная модель языка позволила исправить достаточно много ошибок распознавания, которые не укладывались в модель языка. В таблице 3 показаны примеры исправления таких ошибок.

5. Индивидуализация транскрипций

5.1. Моделирование особенностей произношения дикторов

Для превращения орфографического текста в фонемный был разработан и сформирован набор контекстно-зависимых правил, по которым орфогра-

Таблица 3

**Примеры исправления ошибок распознавания
при помощи биграммной модели языка**

Было сказано	Результат распознавания с использованием свободного порядка слов	Результат распознавания с использованием биграммной грамматики
Доброго ранку	До в в о ранку	Доброго ранку
Шановні народні депутати запрошені та гості верховної ради	Шановні народі депутати запрошені та гості верховної ради	Шановні народні депутати запрошені та гості верховної ради
Прошу вас шановні колеги займати вас свої робочі місця	Прошу в о з шановні колег і й з е мати з в й й робоче місця	Прошу вас шановні колеги займати вас свої робочі місця
Прошу займати робочі місця	Прошу з е мав те й робоче й місця	Прошу займати робочі місця
Прошу підготуватися до реєстрації	Прошу б й готуватися до реєстрації б	Прошу підготуватися до реєстрації

фическое слово превращается в последовательность фонетических символов (путём преобразования одной последовательности символов в другую). Причём генерируется сразу несколько вариантов транскрипции для случаев неоднозначностей, заданных в правилах.

Для всех дикторов был создан общий вариант транскрибирования на основе правил литературного произношения для украинского языка. Кроме этого, все дикторы были разделены на группы, для которых разработаны свои правила индивидуализированного транскрибирования, которые заменяют или дополняют основной вариант.

Результаты изучения речи многих дикторов свидетельствуют, что ни один из них не придерживается орфоэпических правил произношения в полном объёме. В первую очередь это касается запрещённых литературной нормой регрессивной ассимиляции по глухости в паре фонем «звонкая + глухая» и оглушение согласных перед паузой. Дикторы с такими особенностями произношения были выделены в отдельную группу. Обработанный материал свидетельствует, что звонкие согласные в речи таких дикторов в позиции перед глухими оглушаются: (*тобто* → *т о п т о*; *підтримати* → *п' і т т р И м а т и*). Случаи оглушения звонких согласных перед паузой встречаются у большинства дикторов: (*робив* → *р о б И ф*).

Были выделены и многие другие характерные черты произношения разных дикторов: редуцирование окончаний некоторых слов (прилагательных, глаголов) в слитной речи (*шановний* → *ш а н О в н и*; *доброго* → *д О б р о*), «акание» (*робити* → *р а б И т и*), твёрдое произношение мягких согласных (*синього* → *с И н о го*) и др.

Такие тенденции моделируются путём изменения правил перехода от одних последовательностей символов к другим и расширением существующих правил.

В таблице 4 приведены примеры индивидуализированных транскрипций для нескольких слов. В основном, здесь задействованы правила оглушения и редуцирования окон-



чаний в словах. Для некоторых слов (служебных в том числе) задаётся несколько вариантов транскрипций — с ударением на разных слогах (если в языке возможны разные варианты прочтения таких слов) или вообще без ударения (коли → к о л И ; к О л и ; к о л и).

Таблица 4

Примеры модификации транскрипций слов

Слово	Литературная транскрипция	Модифицированная транскрипция
відповідно	в' і д п о в' і д н о	в' і т п о в' і д н о
головуючий	г о л о в У й у ч и й	г о л о в У й у ч и
доброго	д О б р о г о	д О б р о
коли	к о л И	к о л И к О л и к о л и
народного	н а р О д н о г о	н а р О д н о
підтримати	п' і д т р И м а т и	п' і т т р И м а т и п' і т р И м а т и
при	п р И	п р И п р и
робив	р о б И в	р о б И ф
робити	р о б И т и	р а б И т и
синього	с И н' о г о	с И н о г о
тільки	т' І л' к и	т' І л' к и т' і л' к и
тобто	т О б т о	т о п т о
шановний	ш а н О в н и й	ш а н О в н и

5.2. Правила индивидуализированной модификации транскрипций

Все правила индивидуализированной модификации транскрипций можно разделить на несколько групп.

К позиционным изменениям звуков в речевом потоке (изменения, которые зависят от таких условий, как позиция звука в слоге/слове, ударность/безударность, и др. [4]) относим:

- 1) кроме редукции безударных е, и, о (произносятся как е^н, и^е, о^н), также слабое произношение о как а в безударной позиции, реже встречается редукция безударных гласных до полного их исчезновения (тепер → т и п Е р, зозуля → з у з У л' а, боротьба → б а р а д' б А или б р а д' б А);
- 2) оглушение звонких согласных перед паузой (бріd → б р' І т, зараз → з А р а с);
- 3) редукцию в терминальных частях слов в процессе речи — исчезновение согласного звука в окончаниях - ого, -их, -ий, -іх, -ій, -ії, -ої, -еї,

-ою, -сю, -ити и подобных (короткий → к о р О ч ш и, синіх → с И н' і, безпекою → б е с п Е к о у); исчезновение конечного гласного звука в окончаниях -ою, -ею, -сю и подобных (доброю → д О б р о й, землю → з е м л Е й) и др.

К комбинаторным изменениям (качественные и количественные изменения соседних звуков [4]) относим:

- 1) полную регрессивную ассимиляцию по глухости в паре фонем «звонкая + глухая» на стыке любых морфем и на стыке слов (без причины → б е с п р и ч И н и, розсунути → р о с с У н у т и, книжка → к н И ш к а, сядьте → с' А т' т е);
- 2) ассимиляцию по мягкости свистящих и шипящих согласных, губных и заднеязычных согласных (злі → з' л' І, шлях → ш' л' А х, квітка → к' в' І т к а);
- 3) произношение удлинённых и удвоенных согласных звуков как одного короткого звука, произношение двух гласных звуков как одного звука (віддати → в' і д А т и, знання → з н а н' А, зоопарк → з о п А р к, аеропорт → а р о п О р т);
- 4) неполное упрощение сочетаний согласных (чесний → ч Е с т н и й) и др.

5.3. Индивидуализированные транскрипции словарей для групп дикторов

Отбор групп дикторов из обучающей выборки в связи с особенностями произношения проходил в несколько этапов. Сначала были внесены в отдельную группу дикторы с литературным произношением, после этого производилась классификация индивидуальных особенностей речи оставшихся дикторов. В результате мы получили 15 групп дикторов с разными особенностями произнесения, для каждой группы были сгенерированы свои правила транскрибирования словарей (15 типов затранскрибированных словарей).

Для проверки соответствия индивидуализированных транскрипций словарей и живой речи была проведена процедура распознавания речи каждого диктора (вне зависимости от группы, в которую он попал) с использованием всех 15 типов словарей.

После анализа результатов эксперимента каждому диктору были приписаны те правила транскрибирования, которые повысили надёжность распознавания его речи [5]. В результате для всех групп дикторов сформировался свой словарь с индивидуализированными правилами транскрибирования, который повысил надёжность распознавания речи (таблица 5).

Таблица 5

Надёжность распознавания с применением индивидуализированных словарей

Диктор	Длительность записи (сек.)	Надёжность распознавания с использованием лит. словаря (%)	Надёжность распознавания с использованием инд. словаря (%)	Изменение надёжности распознавания (%)
STO	2184	58,61	59,34	0,73
TER	1124	75,18	75,53	0,35
CER	1756	67,14	68,05	0,91
VAS	1629	67,40	69,00	1,6
BAB	1194	58,12	59,72	1,6



Рис. 2. Соотношение литературной и индивидуализированных транскрипций речи дикторов из обучающей выборки

Эксперимент показал, что существуют нелитературные правила, которые могут быть приписаны большинству дикторов, например: редукция окончаний слов, обусловленная слишком быстрым темпом речи или эмоциональностью высказывания; ассимиляция по глухости, обусловленная влиянием русского языка, и др. Для многих дикторов характерно произношение удлинённых и удвоенных звуков как одного короткого звука. Часто встречается произнесение ударных гласных как безударных (в речевом потоке ударение исчезает не только в односложных словах, но и в многосложных). Остальные правила характерны для одного или нескольких дикторов (рисунок 2).

5.4. Генерирование словаря транскрипций для обучающей выборки

Словарь транскрипций для обучающей выборки был сформирован на основе 15 типов индивидуализированных словарей. Каждое слово в словаре имеет то количество транскрипций, сколько существует гипотетических вариантов произнесения этого слова каждым диктором из обучающей выборки. До использования правил индивидуализации словарь транскрипций состоял приблизительно из 36 000 литературных транскрипций слов; после использования этих правил, а также комбинаций правил словарь увеличился до 285 000 транскрипций слов.

6. Результаты экспериментов по распознаванию слитной речи

Эксперименты проводились на описанной контрольной выборке в виде записей заседаний в течение одного дня. Надёжность распознавания сильно отличается в зависимости от того, какие дикторы попали в выборку.

Таблица 6 представляет результаты распознавания для некоторых дикторов, где также приведены длина обучающей выборки и темп произнесения для каждого диктора. Анализ результатов показывает, что в среднем указанные факторы (длина обучающей выборки и темп речи) влияют на надёжность распознавания.

Надёжность распознавания для отдельных дикторов сильно отличается — от 50% до 90%. В среднем по всей контрольной выборке она составляет 72.3%. Использование индивидуализированных транскрипций улучшило надёжность распознавания в среднем на 1,5%.

Таблица 6

Надёжность распознавания для некоторых дикторов

Диктор	Длина обучающей выборки (сек.)	Длина контрольной выборки (сек.)	Количество слов в контрольной выборке	Темп (слов/сек.)	Надёжность распознавания (%)
LIT	15 805	2336	5 721	2,45	79,85
POR	3 715	411	853	2,08	80,30
MOR	1 728	362	950	2,62	70,74
SIM	1 484	125	255	2,04	80,00
MAT	1 305	174	292	1,68	80,14
KLU	998	107	209	1,95	86,60
KIN	585	223	417	1,87	64,27
ONI	483	100	209	2,09	79,90
MIS	195	148	312	2,11	69,87
ZVA	25	101	205	2,03	80,00
GOL	0	379	790	2,08	78,48
KAP	0	375	927	2,47	80,91

Время распознавания для компьютера Pentium 2GHz составляет около 10 секунд для одной секунды речи. Применение алгоритмов ускорения принятия решений [6] позволит достичь распознавания речи в реальном времени.

Заключение

Статья описывает экспериментальную систему автоматизированного стенографирования.

Показана возможность построения таких систем при условии повышения надёжности распознавания речи до необходимых для практических применений показателей.

Предложено использовать индивидуальную информацию о дикторах для улучшения надёжности распознавания.



Литература

1. V. Pylypenko. Extra Large Vocabulary Continuous Speech Recognition Algorithm based on Information Retrieval // Proc. of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007). Antwerp, Belgium, 2007.
2. S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, P. Woodland. The HTK Book. Cambridge University Engineering Department, 2002.
3. Винцюк Т.К. Анализ, распознавание и смысловая интерпретация речевых сигналов. Киев: Наукова думка, 1987. 264 с.
4. Сучасна українська літературна мова. Фонетика: Навч. посібник для студентів-філологів. К.: Видавничо-поліграфічний центр «Київський університет», 2002. С. 60.
5. V. Pylypenko, V. Robeiko. Experimental System of Computerized Stenographer for Ukrainian Speech // Proc. of the 13th International Conference «Speech and Computer: SPECOM'2009». St.-Petersburg, Russia, 2009. P. 67—70.
6. Пилипенко В.В. Распознавание дискретной и слитной речи из сверхбольших словарей на основе выборки информации из баз данных // Искусственный интеллект. № 3. 2006. С. 548—557.

Пилипенко Валерий Васильевич

Старший научный сотрудник Международного научно-учебного центра информационных технологий и систем (г. Киев, Украина). Окончил Московский физико-технологический институт в 1984 году. Специалист в области распознавания речи.

Робейко Валентина Васильевна

Окончила Институт филологии Киевского национального университета им. Т. Шевченко. Специалист в области современного украинского языка. Работает младшим научным сотрудником в Международном научно-учебном центре информационных технологий и систем (г. Киев, Украина), аспирантка. Область интересов: синтез и распознавание речи, фонетика.

Применение метода нелинейного отображения многомерных данных в задаче постановки правильного произношения звуков в составе отдельных слов



*Губочкин И.В.,
аспирант*

В статье предложен метод обучения произношению звуков в составе коротких слов и фраз с интерактивной визуализацией результатов. Его основу составляет совместное использование аппарата скрытых марковских моделей и нелинейного отображения многомерных данных. Приведён пример практического применения метода в задаче обучения произношению некоторых фонем в составе слов английского языка.

In the article the method of sounds pronunciation training in short words and phrases with interactive results visualization is suggested. Its foundation is the joint use of hidden Markov model approach and the multidimensional data nonlinear mapping. An example of method's practical application in the pronunciation training task for certain phonemes in English words is presented.

Введение

Системы компьютерного обучения языку (КОЯ) в настоящее время получают всё большее распространение. В этих системах компьютер обеспечивает немедленную реакцию на действия обучающегося и позволяет ему самостоятельно выбирать скорость изучения языка. Таким образом, реализация эффективной обратной связи между компьютером и человеком является важнейшей задачей при построении системы КОЯ. Для её решения разработано множество методов и подходов [1—4]. Среди них большой класс составляют методы, основанные на сравнении входного сигнала с некоторым эталоном в частотной или временной области. Основной недостаток систем, построенных по этому принципу, состоит в том, что даже при очень хорошем произношении входной сигнал и эталон могут иметь совершенно разные спектры или формы во времени [5].

Кроме того, получаемые результаты достаточно трудно интерпретировать, поскольку нет простого соответствия между артикуляционными движениями и отображаемыми результатами.

Описанные выше проблемы создают трудности при обучении языку с использованием средств автоматизированного контроля. Поэтому важно, чтобы ответ системы КОЯ был максимально понятным пользователю. Кроме того, обучающемуся необходимо знать не только тот факт, что он совершает ошибку, но также и её тип. Это позволит ему самостоятельно установить, что именно он делает неправильно. В [6] был предложен метод обучения произношению отдельных фонем с использованием нелинейного отображения данных Сэммона. Этот метод позволяет проводить сравнение произношения с множеством эталонных реализаций и диагностику наиболее распространённых ошибок. Представляемые результаты могут быть наглядно отображены на экране монитора компьютера. В настоящей работе даётся обобщение этого метода на случай тренировки произношения звуков внутри слов и коротких фраз.

Алгоритм отображения

Алгоритм отображения речевых сигналов состоит из двух этапов. На первом этапе происходит сегментирование входного сигнала на фонемы, соответствующие изучаемому слову, а на втором — отображение нужной фонемы.

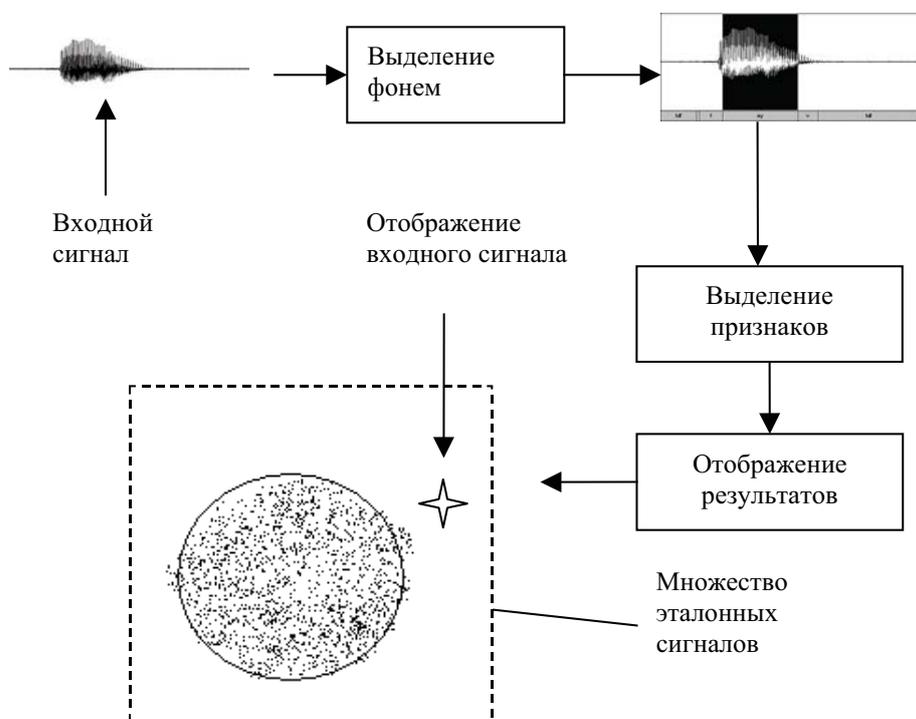


Рис. 1. Схема работы алгоритма

Для осуществления первого этапа наиболее перспективным представляется использование аппарата скрытых марковских моделей (СММ) [7]. После предварительной тренировки моделей отдельным фонем на обучающей выборке они составляются в слова или короткие фразы. Полученные составные модели в дальнейшем используются для определения, при помощи алгоритма Витерби, местоположения границ отдельных фонем. Таким образом, в результате выполнения первого этапа алгоритма мы получаем фонетическую разметку входного сигнала.

Рассмотрим теперь второй этап. Выделив во входном сигнале интервал, соответствующий интересующей нас фонеме, мы можем вычислить её вектор признаков и произвести его отображение на плоскости относительно множества эталонных сигналов. На рисунке 1 представлена схема работы алгоритма.

Если задача разметки речевого сигнала на фонемы при известной транскрипции и вычисления векторов признаков в настоящее время в основном решена, то в части представления результатов до сих пор ведутся интенсивные исследования. Одним из перспективных подходов в этом направлении следует признать метод нелинейного отображения многомерных данных, предложенный Сэммоном [8], который заключается в следующем.

Обозначим через δ_{ij} расстояние между векторами x_i и x_j исходного пространства 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:

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

где N — число векторов множества входных данных X .

Для того чтобы получить оптимальное в смысле сохранения расстояния отображение, найдём минимум E (1) по y_j для случая использования в качестве δ_{ij} евклидова расстояния. Метод наискорейшего спуска приводит к следующему рекуррентному уравнению:

$$y_i(l+1) = y_i(l) - \alpha \frac{\partial E}{\partial y_j} = y_i(l) + \left[2\alpha / \left(\sum_{i=1}^{N-1} \sum_{j=i+1}^N \delta_{ij} \right) \right] \sum_{\substack{j=1 \\ j \neq i}}^N [(\delta_{ij} - d_{ij}) / (\delta_{ij} d_{ij})] \cdot [y_i(l) - y_j(l)]. \quad (2)$$

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

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

где ε — некоторая константа, $\varepsilon > 0$. На практике значения ε обычно устанавливаются в интервале $10^{-4} \dots 10^{-9}$.

Для случая добавления в исходное множество X ещё одного вектора данных воспользуемся следующей процедурой [6]. Обозначим через Δ матрицу расстояний между векторами множества X :

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

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 будут определяться как

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

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

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

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

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

Останов работы алгоритма (7) происходит при выполнении условия (3).

Программа экспериментальных исследований

Для экспериментальных исследований было сформировано две группы дикторов по четыре человека каждая. Первую группу составляли дикторы, свободно владеющие английским языком. Во вторую группу были включены студенты 1—5 курсов языкового вуза. В качестве тестового был выбран следующий набор слов английского языка: bit [blt], boot [bu:t], bottom ['bɒtəm], she

[ʃi:], way [weɪ], winner [ˈwɪnə]. В этих словах встречаются фонемы, относящиеся к различным группам: фрикативы (/SH iy/), гласные (/b IH tcl t/, /b UW tcl t/), носовые (/b aa tcl t EM/, /w ih NX axr/), полугласные (/W ey/). Транскрипции слов в данном случае приведены в соответствии с обозначениями фонем в широко известной речевой базе ТИМІТ [9]. Каждое слово проговаривалось четыре раза дикторами из обеих групп. Для ввода речевого сигнала в ПК 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 на уровне 16 кГц — это общепринятая частота при обработке устной речи.

Реализация предложенного метода делится на три независимых во времени этапа. На первом этапе формируется база априорных данных по каждой реализации всех фонем. Для этого вычисляется вектор признаков, описывающий ту или иную реализацию. В [6] было показано, что наиболее предпочтительными для использования являются кепстральные коэффициенты линейного предсказания с неравномерным частотным разрешением, которые, в свою очередь, рассчитываются на основе коэффициентов линейного предсказания с неравномерным частотным разрешением [10]. Несмотря на то, что по своим свойствам они близки к широко известным мел-кепстральным коэффициентам (MFCC), их использование в задаче описания характеристик фонемы обусловлено прежде всего удобством практического применения.

Для получения коэффициентов линейного предсказания с неравномерным частотным разрешением сначала найдём коэффициенты автокорреляции r входного сигнала x .

$$\begin{aligned} W^{(0)} &= x \\ r(0) &= x^T x \\ W^{(m)}(n) &= -\lambda W^{(m-1)}(n) + W^{(m-1)}(n-1) + \lambda W^{(m)}(n-1), \quad 1 \leq n \leq N \\ r(m) &= x^T W \\ m &= \overline{1, p} \end{aligned} \quad (8)$$

Здесь N — объём выборки, p — порядок модели, λ — коэффициент деформации, определяемый как

$$\lambda \approx 1,0674 \left(\frac{2}{\pi} \tan^{-1}(0,06583 f_s / 1000) \right)^{1/2} - 0,1916, \quad (9)$$

где f_s — частота дискретизации в Гц.

В дальнейшем непосредственно для получения коэффициентов линейного предсказания с неравномерным частотным разрешением а воспользуемся автокорреляционным методом [11]:

$$\begin{aligned} E^{(0)} &= r(0) \\ k_i &= \left\{ r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(i-j) \right\} / E^{(i-1)}, \quad 1 \leq i \leq p \\ \alpha_i^{(i)} &= k_i \\ \alpha_j^{(i)} &= \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \\ E^{(i)} &= (1 - k_i^2) E^{(i-1)} \\ \alpha_m &= \alpha_m^{(p)}, \quad 1 \leq m \leq p \end{aligned} \quad (10)$$

Кепстральные коэффициенты вычисляются по следующей формуле [7]:

$$\begin{aligned}c_m &= a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \leq m \leq p \\c_m &= \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad m > p\end{aligned}\tag{11}$$

В нашем эксперименте векторы признаков рассчитывались по реализациям фонем, взятым из речевой базы TIMIT, диалект DR1 (New England). Поскольку количество реализаций каждой фонемы составляло от сотни до нескольких тысяч, то векторы признаков предварительно подвергались векторному квантованию по методу k гармонических средних [12]. Выбор данного метода обусловлен тем, что он даёт меньшую ошибку квантования и быстрее сходится по сравнению с классическим методом k -средних. После векторного квантования производится отображение полученных результатов на двумерную плоскость по алгоритму (1)—(3), а в качестве меры расстояния δ_{ij} между векторами кодовой книги используется евклидова метрика.

На втором этапе производится оценка параметров монофонных СММ для реализаций фонем из обучающего подмножества TIMIT при помощи специализированного пакета НТК (Hidden Markov Model Toolkit) [13]. В качестве вектора признаков использовались 12 мел-кепстральных коэффициентов и логарифм энергии, а также их первая и вторая производные по времени (всего 39 элементов). Для всех фонем, кроме модели тишины, прототипы СММ имели 3 состояния, 39-элементный вектор признаков и 3 гауссовых смеси. Модель тишины имела одно состояние.

Начальное оценивание параметров моделей осуществлялось при помощи алгоритма Витерби. После этого была проведена их переоценка с использованием алгоритма Баума-Уэлча [7].

На третьем этапе обучающийся последовательно произносит реализации определённого слова. Каждая такая реализация сначала сегментируется на фонемы при помощи программы HVITE из пакета НТК. Затем по фрагменту речевого сигнала, который соответствует интересующей нас фонеме, вычисляется вектор признаков. Полученный вектор отображается на двумерную плоскость при помощи алгоритма (6)—(7). Таким образом, на плоскости кроме множества точек, соответствующих произношению фонемы, появляется ещё одна точка, которая характеризует расположение входного сигнала относительно данного множества. Учитывая это, диктор может корректировать своё собственное произношение, приближая его к эталонному.

Результаты эксперимента

В качестве параметров алгоритма визуализации были установлены следующие:

- порядок АР-модели $p = 8$;
- число кепстральных коэффициентов $m = 12$;

- параметр алгоритма отображения $\alpha = 0,4$;
- относительная ошибка отображения $\varepsilon = 10^{-5}$;
- максимальное количество векторов кодовой книги $k = 100$.

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

Для примера рассмотрим процесс работы алгоритма для слова «way», произнесённого одним из дикторов первой группы. На рисунке 2 показан результат определения границ фонем для одной из реализаций.

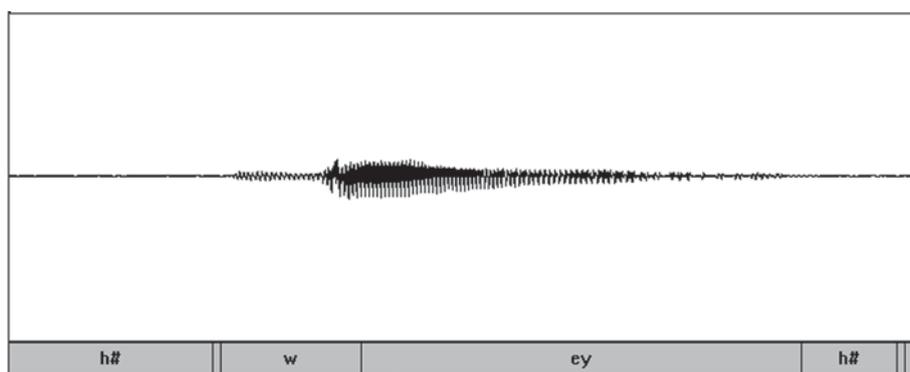


Рис. 2. Найденные границы фонем для слова «way»

После выделения границ фонем из входного сигнала выбирается фрагмент, соответствующий обучаемой фонеме. По нему рассчитывается вектор кепстральных коэффициентов с неравномерным частотным разрешением. На рисунке 3 на верхнем графике показан сигнал, соответствующий слову «way», на котором обозначены найденные границы фонемы /w/. На нижнем графике показаны значения коэффициентов вектора признаков (11), рассчитанных по выделенному участку сигнала.

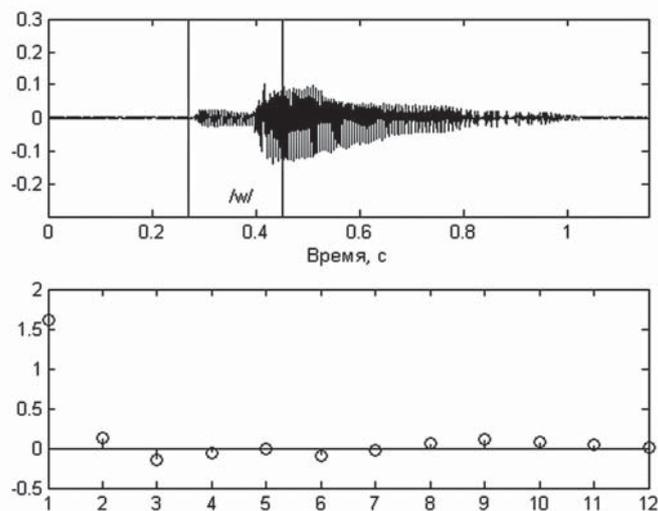


Рис. 3. Результаты выделения границ фонемы /w/

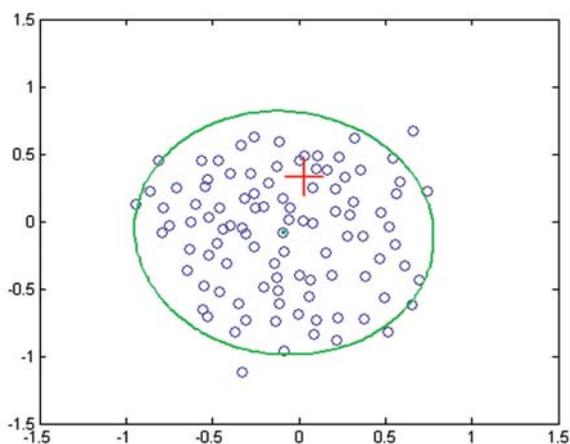


Рис. 4. Отображение реализации фонемы /w/ от диктора из первой группы

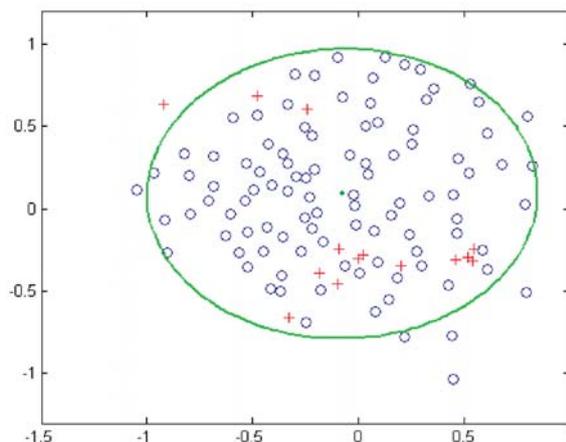


Рис. 5. Отображение реализаций эталонных фонем и фонем дикторов первой группы

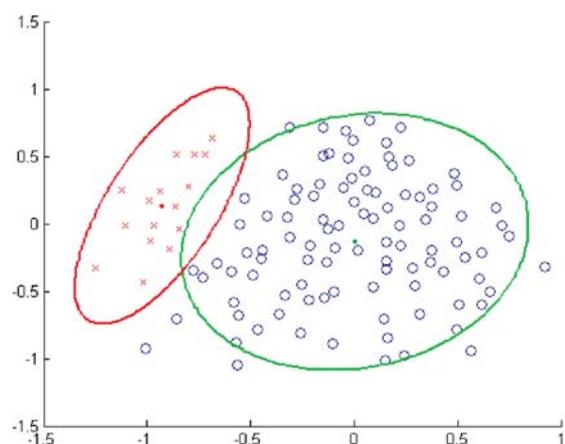


Рис. 6. Отображение реализаций эталонных фонем и фонем дикторов второй группы

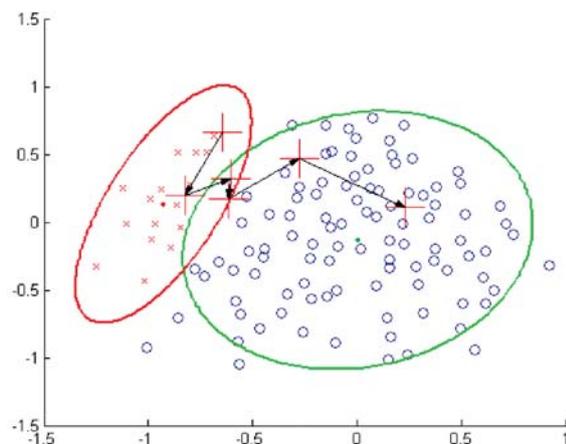


Рис. 7. Результаты обучения

На рисунке 4 показан результат отображения в двумерное пространство полученного вектора признаков при помощи алгоритма (6)—(7).

Здесь символами 'o' обозначены эталонные реализации фонем, а символом '+' обозначено положение вектора признаков входного сигнала. Эллипс ограничивает область эталонных сигналов. Поскольку диктор имеет хорошо поставленное английское произношение, произнесённая им фонема сразу же попала в область эталонных значений. Для подтверждения полученного результата на рисунке 5 показано отображение всех реализаций фонемы /w/ для дикторов первой группы.

Нетрудно видеть, что почти все из них попали в область эталонных значений. Аналогичные результаты получены и для остальных фонем из тестового

набора. Это позволяет говорить о том, что разработанный алгоритм является инвариантным к выбору эталонных сигналов.

Рассмотрим теперь результаты работу алгоритма для дикторов из второй группы, представленные на рисунке 6.

Здесь символами 'x' дополнительно показаны реализации фонемы /w/ для дикторов из второй группы. Видно, что они образуют компактный кластер (его границы показаны красной линией) вблизи границы области эталонов. Отсюда можно сделать вывод, что произношение дикторов, входящих во вторую группу, всё же существенно отличается от произношения носителей языка. Задачей обучаемого, таким образом, является изменить собственное произношение, чтобы переместить 'x' в область эталонных сигналов. На рисунке 7 показан пример такого обучения.

Из рисунка видно, что чем более чётко проговаривается фонема /w/ внутри слова, тем ближе к эталонному сигналу (и дальше от области неправильного произношения) располагаются её реализации.

В рассматриваемом эксперименте не проводилась какая-либо классификация ошибок произношения. Однако при наличии такой информации предложенный алгоритм легко обобщить на данный случай. Для этого на множестве точек двумерного отображения необходимо выделить области, соответствующие какой-либо типичной ошибке (аналогично рисунку 6). Рассматриваемый случай схематично показан на рисунке 8.

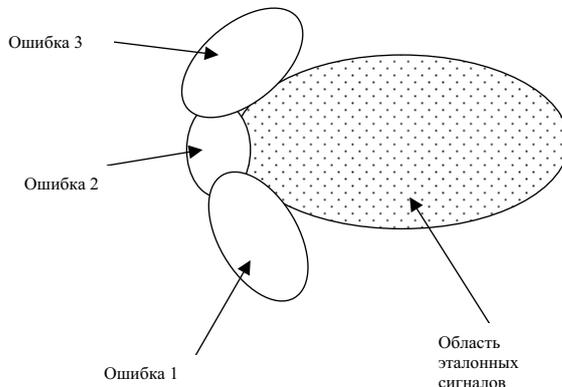


Рис. 8. Иллюстрация обобщения предложенного алгоритма на случай наличия нескольких типов ошибок произношения фонемы

Таким образом, попадая в ту или иную область, обучаемый может самостоятельно диагностировать проблемы в собственном произношении и при необходимости производить его коррекцию.

Численная оценка качества произношения

Рассмотренные выше результаты оценки произношения дают в первую очередь качественные результаты. Для более точного определения правильности произношения следует



задать некоторую его количественную оценку. Из практических соображений следует, что такая оценка должна обладать следующими двумя свойствами:

- все возможные значения оценки должны находиться в некоторой ограниченной области (например, на интервале $[0..1]$);
- принимать малые значения, в случае если получаемый результат выходит из области эталонных значений или же находится на её периферии;
- принимать значения, близкие или равные максимуму, в случае если получаемый результат находится в глубине области эталонных значений или близок к её центру.

Можно предложить множество способов определения оценки, обладающей указанными выше свойствами. Один из наиболее простых способов сделать это — использовать функцию распределения вероятностей расположения координат точек, полученных в результате отображения эталонных сигналов на двумерную плоскость. Таким образом, оценка качества произношения определяется следующим образом:

$$\rho(Y) = 1 - [P(M+|M-Y) - P(M-|M-Y)]. \quad (12)$$

Здесь P — функция распределения вероятностей расположения координат точек отображения, M — математическое ожидание, Y — координаты отображения входного сигнала.

Для возможности практического применения приведённой выше оценки (12) требуется задать функцию плотности вероятности $\rho(Y)$. Такая функция должна быть определена на всём множестве значений Y и иметь симметричную форму относительно значения математического ожидания M . Подобным требованиям удовлетворяет широко известная функция плотности вероятности нормального распределения в её двумерном варианте. Для подтверждения корректности применения нормального закона при описании распределения значений координат точек отображения был проведён следующий эксперимент.

Для каждой фонемы из тестового набора было произведено отображение её реализаций на двумерную плоскость с помощью алгоритма (1)—(5). Затем каждое полученное множество точек было проверено на соответствие нормальному закону с помощью универсального критерия Дурника-Хансена (Doornik-Hansen) [14] на уровне значимости 0,05. Полученные результаты приведены в таблице 1.

Таблица 1

Значения решающей статистики Дурника-Хансена

Фонема	/SH/	/IH/	/UW/	/EM/	/NX/	/W/
E_p	0,3029	0,7013	0,8174	0,9678	0,3726	0,1843

Здесь E_p — значения решающей статистики, используемой в критерии.

Из таблицы видно, что, по крайней мере, на выбранном уровне значимости тест был пройден для всех фонем (значения решающей статистики E_p превы-

шают уровень значимости). Данный результат позволяет сделать вывод о корректности использования нормального закона для описания распределения координат точек отображения.

Вспользуемся теперь оценкой (12) для определения качества произношения дикторов из первой и второй тестовых групп. Для этого по множеству координат точек отображения эталонных реализаций выбранной фонемы r производилось оценивание математического ожидания $\bar{\mathbf{M}}_r$ и ковариационной матрицы $\bar{\Sigma}_r$:

$$\begin{aligned}\bar{\mathbf{M}}_r &= \frac{1}{N_r} \sum_{i=1}^{N_r} \mathbf{y}_{i,r}, \\ \bar{\Sigma}_r &= \frac{1}{N_r - 1} \sum_{i=1}^{N_r} (\mathbf{y}_{i,r} - \bar{\mathbf{M}}_r)^T (\mathbf{y}_{i,r} - \bar{\mathbf{M}}_r),\end{aligned}\tag{13}$$

где N_r — число реализаций r -й фонемы, $\mathbf{y}_{i,r}$ — вектор координат i -й точки отображения. Затем по реализациям фонем, полученных от дикторов из тестовых групп, рассчитывалась оценка $\rho(Y)$. Полученные значения, усреднённые по всем дикторам первой и второй групп, показаны в таблицах 2 и 3 соответственно.

Таблица 2

Оценка качества произношения дикторов первой группы

Фонема	/SH/	/H/	/UW/	/EM/	/NX/	/W/
$\rho(Y)$	0,2736	0,2707	0,2775	0,4194	0,3010	0,3168

Таблица 3

Оценка качества произношения дикторов второй группы

Фонема	/SH/	/H/	/UW/	/EM/	/NX/	/W/
$\rho(Y)$	0,3662	0,1776	0,2968	0,3258	0,2976	0,2902

Представленные в таблицах результаты наглядно свидетельствуют о том, что дикторы с хорошо поставленным английским произношением получают более высокие оценки. Недостаточно сильная разница в значениях $\rho(Y)$ (и даже превышение усреднённого качества произношения дикторов второй группы над результатами дикторов первой группы для фонем /SH/ и /UW/), по-видимому, обусловлена недостаточными размерами выборок и большой вариативностью результатов внутри каждой группы.

Выводы

Полученные в ходе экспериментальных исследований результаты позволяют сделать вывод об эффективности разработанного алгоритма в решении задачи обучения произношению. Кроме того, в метод легко может быть добавлена возможность диагностики наиболее часто встречающихся ошибок путём добавления во множество априорных



данных реализаций фонем, соответствующих какой-либо типичной ошибке произношения. Зная порядковые номера таких реализаций, на двумерном отображении легко можно выделить области, соответствующие различным классам ошибок (см. рисунок 8).

Ещё одно преимущество представленного алгоритма состоит в том, что, в случае если предварительно определён язык (и/или диалект), произносительные транскрипции могут быть определены независимо от диктора. Эта возможность является особенно актуальной для практического применения в реальном учебном процессе.

Также предложена численная оценка качества произношения и рассмотрена методика её вычисления.

Дальнейшее развитие метода может быть направлено на повышение точности диагностики ошибок. Здесь может быть выделено два пути. Первый — учёт ошибок, связанных с неправильной длительностью произношения отдельных фонем и ударениями, поскольку предложенный алгоритм анализирует, в основном, различия в спектральной области. А второй путь — это совместное использование акустических признаков и информации, получаемой из видеоизображения обучаемого [15].

Литература

1. *Akahane-Yamada R., McDermott E.* Computer-based second language production learning by using spectrographic representation and HMM-based speech recognition scores. Proceedings of ICSLP, Sydney, Australia 1998.
2. *Nouza J.* Training speech through visual feedback patterns. Proceedings of ICSLP Sydney, Australia 1998.
3. *Molholt G.* Computer-assisted instruction in pronunciation for Chinese speakers of American English. TESOL Quarterly 22, 91—111, 1988.
4. *Tsubota Y., Kawahara T., and Dantsuji M.* CALL system for Japanese students of English using pronunciation error prediction and formant structure estimation. In InSTIL 2002 Advanced Workshop, 2002.
5. *Neri A., Cucchiari C., Strik H.* Feedback in Computer Assisted Pronunciation Training: technology push or demand pull? Proceedings of ICSLP 2002, Denver, USA, P. 1209—1212.
6. *Губочкин И.В.* Алгоритм визуализации речевых сигналов для интерактивного обучения правильному произношению // Речевые технологии/ 2008. №3, С. 72—80,
7. *Rabiner L.R., Juang B.-H.* Fundamentals of speech recognition. Prentice Hall, Englewood Cliffs, NJ, 1993.
8. *Sammon J.W.* A non-linear mapping algorithm for data structure analysis. *IEEE Trans. Computers*, CC-18(5):401—409. 1969.
9. *William M.Fisher, George R.Doddington, and Kathleen M.Goudie-Marshall.* «The DARPA Speech Recognition Research Database: Specifications and Status», Proceedings of DARPA Workshop on Speech Recognition. P. 93—99, Feb. 1986.
10. *Härmä 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.

11. Марпл С.Л.-мл. Цифровой спектральный анализ и его приложения. М.: Мир, 1990. 584 с.
12. Bin Zhang. «Generalized K-Harmonic Means — Boosting in Unsupervised Learning», Hewlett-Packard Labs, Technical Report HPL-2000-137, 2000.
13. Young *et al.*, Hidden Markov Model Toolkit v3.1 reference manual, Technical report, Speech group, Cambridge University Engineering Department, December. 2001.
14. Doornik J.A. and Hansen H. An omnibus test for univariate and multivariate normality. Discussion Paper W4&91. Nuffield College, Oxford, UK, 1994.
15. Vicsi K. An Overview of Speech Training Methods Based on Multi Modal Feedback. IOS Press, Amsterdam. P. 1—10, 2004.
16. Фукунага К. Введение в статистическую теорию распознавания образов: Пер. с англ. М.: Наука, 1979. 368 с.

Губочкин Иван Вадимович

Инженер-программист ООО «МФИ-Софт», аспирант кафедры математики и информатики Нижегородского государственного лингвистического университета им. Н.А. Добролюбова. Область научных интересов — автоматическая обработка речевых сигналов. Автор девяти научных работ.
E-mail: jhng@yandex.ru.



Автоматическое определение изменений эмоционального состояния по речевому сигналу

*Лукьяница А.А.,
кандидат физико-математических наук*

*Шишкин А.Г.,
кандидат физико-математических наук*

Настоящая работа посвящена проблеме создания автоматической компьютерной системы, позволяющей определить изменение эмоционального состояния диктора на основе речевого сигнала. Системы подобного типа относятся к системам распознавания образов, которые обычно состоят из двух основных частей: первая должна выделять из речи наиболее информативные характерные признаки, а вторая (классификатор) — на основе выделенных признаков принимать решение об изменениях в эмоциональном состоянии человека. В данной работе речевой сигнал описывается 211 характерными признаками, что позволило в итоге получить достоверность распознавания состояния диктора с точностью 97.2%. Классификатор был построен на основе метода опорных векторов. Проведённые исследования показали, что число признаков может быть сокращено до 57 самых важных, что приводит к снижению точности лишь на 1.1%.

This paper reports results in development of a computer system that explores the emotional state of a human by his speech. The system consists of two parts: speech features extraction and human state classification. We classify speech into two emotional states based on 211 features with total classification error less than 3%. The classification is performed by support vector machine method. We show that the number of speech features can be reduced to 57 most important ones, thus increasing classification error by 1.1% only.

Ключевые слова: эмоциональное состояние, частота основного тона, форманты, кепстральные коэффициенты, линейное предсказание, метод опорных векторов

1. Введение

При изменении эмоционального состояния в человеческом организме происходят сложные процессы, которые в конечном итоге находят отражение в виде мышечных сокращений, в том числе и в голосовом тракте. Это даёт возможность бесконтактного определения эмоционального состояния человека по изменениям в системе речеобразования. Системы подобного типа, как правило, состоят из двух основных частей, первая из которых должна выделять из речи наиболее информативные характерные признаки, а вторая — на основе выделенных признаков принимать решение, является ли данный образец речи спокойным или соответствует стрессовым изменениям в состоянии человека. Первую часть будем называть системой выделения характерных признаков, а вторую — распознающей, или классифицирующей, системой.

Многочисленные исследования [1—7] показали, что в состоянии даже лёгкого волнения у человека меняется частота основного тона и нескольких первых формант, изменяется спектральный состав речи, повышается энергия высокочастотных компонент, увеличиваются громкость и темп речи, появляется вибрация, растягиваются гласные, а также происходят другие изменения, которые могут быть описаны в математической форме. В качестве характерных признаков наиболее часто используются следующие величины: F_0 — частота основного тона, а также её дисперсия σ_F^2 и вариабельность δ_F (т.е. слабые изменения); F_1 , F_2 и F_3 — частота первых трёх формант; I — интенсивность речи и её вариабельность δ_I ; J — вибрация (дрожание) голоса; P_I — расположение пиков в звуковой волне; TEO (Teager Energy Operator) [3, 4].

Перечисленные характерные признаки использовались различными исследователями для определения психоэмоционального состояния человека и, судя по описанным результатам, «зарекомендовали себя с положительной стороны». Каждый из этих признаков может изменяться в диапазоне от 1% до 10% при изменении состояния испытуемого, и в совокупности они могут позволить построить систему с высоким уровнем достоверности. Однако это не исключает использования других признаков, которые могут оказаться эффективными именно в российских условиях, с учётом особенностей русской речи.

Настоящая работа организована следующим образом. В следующем разделе описаны выбранные нами характерные признаки и методы их нахождения. Сначала описывается техника отделения речи от пауз, а затем рассматриваются способы вычисления признаков, основанных на определении частоты основного тона, значениях трёх первых формант, а также на вычислении кепстра. Далее приведено краткое описание классифицирующей системы, основанной на методе опорных векторов. В последнем разделе описываются результаты численных экспериментов, направленных на сокращение числа признаков. В заключении сформулированы основные результаты, полученные в настоящей работе.

2. Выделение характерных признаков

Для выделения характерных признаков речевого сигнала на первом этапе его обработки необходимо отделить речь от пауз. При этом осуществляющие указанную операцию методы должны работать в реальном времени. Следовательно, необходимо использовать такие характеристики речевого сигнала, которые являются довольно простыми, но в то же время позволяют надёжно находить начало и конец слова.



2.1. Определение частоты основного тона

Одним из основных характерных признаков, использующихся практически при всех видах анализа речевого сигнала, является частота основного тона и различные производные от неё параметры. Частота основного тона — это та частота, с которой колеблются голосовые связки человека во время произнесения вокализованных звуков. Вследствие того, что данная величина играет чрезвычайно важную роль при получении окончательного результата разрабатываемой системы, необходимо использовать метод, позволяющий определять её с высокой степенью точности. К настоящему моменту существует большое число различных методов вычисления частоты основного тона [8—11]. Наиболее распространёнными являются методы, основанные на использовании автокорреляционной функции и функции нормированной перекрёстной корреляции. Однако простое использование указанных методов приводит во многих случаях к неправильным результатам (например, удвоению истинной частоты основного тона или получению ненулевых значений для невокализованных фрагментов речи). Поэтому нами был использован алгоритм оптимального выбора значения частоты основного тона из имеющихся возможных значений, основанный на использовании метода динамического программирования. Это позволило избежать огромного числа вычислений, неизбежно возникающих при простом переборе всех существующих временных траекторий частоты основного тона, и получать надёжные результаты.

Обозначим речевой сигнал, полученный с частотой дискретизации F_s , как $x(n)$, где $n = 0, 1, 2, \dots$. Тогда нормированная функция перекрёстной корреляции определяется следующим выражением:

$$\phi_i(k) = \frac{\sum_{j=m}^{m+n-1} x(j)x(j+k)}{\sqrt{E(m)E(m+k)}}, \quad k = 0, \dots, K-1; \quad m = iz; \quad i = 0, \dots, M-1, \quad (1)$$

где K — максимальное значение задержки k , i — индекс очередного речевого сегмента (окна), M — число таких сегментов, $z = tF_s$, t — длина сегмента, E — энергия сигнала;

$$E(m) = \sum_{l=m}^{m+n-1} x^2(l). \quad (2)$$

Величина ϕ всегда принадлежит отрезку $[-1; 1]$. При этом $\phi_i(k)$ близка к 1 для задержек k , кратных истинному значению периода основного тона, вне зависимости от наличия или отсутствия быстрых изменений сигнала x .

При наличии низкочастотного шума будут получаться большие значения корреляции для всех задержек в искомом диапазоне. Это будет приводить, в том числе, к тому, что невокализованные фрагменты речевой волны будут классифицироваться как вокализованные. Для устранения данной проблемы будем вычитать из значений сигнала его среднее значение в рассматриваемом окне i , т.е.

$$s_i(j) = x(m+j) - \mu_i, \quad m = iz; \quad j = 0, \dots, n+K-1, \quad (3)$$

где

$$\mu_i = \frac{1}{n} \sum_{j=m}^{m+n-1} x(j). \quad (4)$$

Так как число точек в окне n и максимальная задержка K пропорциональны частоте дискретизации F_s , число необходимых арифметических действий для вычисления $\phi_i(k)$ пропорционально F_s^2 . Для определения частоты основного тона приходится много раз вычислять нормированную функцию перекрёстной корреляции, что даже при использовании быстрого преобразования Фурье приводит к неприемлемо большим затратам вычислительных ресурсов. Поэтому воспользуемся состоящей из двух этапов процедурой, число действий в которой составляет $O(F_s)$. Сначала осуществим прореживание сигнала с частотой F_{ds} , равной

$$F_{ds} = \left\lceil \frac{F_s}{4F_{0\max}} \right\rceil, \quad (5)$$

где $\lceil \cdot \rceil$ обозначает округление до ближайшего целого, а $F_{0\max}$ — максимально возможное значение частоты основного тона.

На первом этапе нормированная функция перекрёстной корреляции для прореженного сигнала вычисляется для задержек $F_{ds}/F_{0\max} \leq k \leq K$. При этом находится максимальное значение ϕ_{\max} . Все значения $\phi_i(k)$, превышающие $\delta_1 \cdot \phi_{\max}$, запоминаются в качестве возможных кандидатов для определения частоты основного тона. Для более точной оценки положения локальных максимумов корреляционной функции и их амплитуд используется параболическая интерполяция по трём значениям, определяющим пики при частоте дискретизации F_{ds} . Если число таких пиков превышает N_1 , то они упорядочиваются в порядке убывания, а потом отбираются первые N_1 пиков.

На втором этапе $\phi_i(k)$ вычисляется с использованием исходного речевого сигнала, но только для семи значений задержки в окрестности каждого из отобранных на первом этапе локальных максимумов корреляционной функции. Аналогичным образом определяется значение ϕ_{\max} и все значения $\phi_i(k)$, превышающие $\delta_1 \cdot \phi_{\max}$. Затем они опять упорядочиваются в порядке убывания, отбираются первые N_1 пиков — и наконец осуществляется параболическая интерполяция для уточнения положения и амплитуды локальных максимумов.

После того как в каждом окне найдены возможные значения частоты основного тона, необходимо отобрать из них одно единственно верное. Для этого воспользуемся методом динамического программирования, предложенным Р. Беллманом [12] для эффективного решения оптимизационных задач. Обозначим через I_i число состояний в окне i ($1 \leq I_i \leq N_1 + 1$). Для каждого окна это число будет равно количеству возможных значений частоты основного тона (вокализованные состояния) плюс одно невокализованное состояние. Пусть C_{ij} — это значение j -го локального максимума ϕ в окне i , а $L_{i,j}$ — соответствующее ему значение задержки.

Определим стоимость назначения окну i вокализованного состояния с периодом $L_{i,j}/F_s$ следующим образом:

$$d_{i,j} = 1 - C_{i,j}(1 - \beta L_{i,j}), \quad 1 \leq j \leq I_i, \quad (6)$$



а стоимость назначения окну i невокализованного состояния — как

$$d_{i,j} = d_0 + \max_j C_{i,j}, \quad (7)$$

где

$$\beta = \frac{\beta_0}{(F_s/F_{0\max})}. \quad (8)$$

Величина β_0 позволяет уменьшать значимость больших задержек, чтобы предпочтительным оказывался выбор задержек с меньшим значением. Очевидно, что такой выбор функций стоимости даёт преимущество значениям $C_{i,j}$, близким к единице, и небольшим величинам задержки для вокализованных сегментов, а для невокализованных — значениям $C_{i,j}$, близким к нулю. Параметр d_0 контролирует вероятность назначения окну вокализованного состояния.

Стоимость перехода Δ от состояния j в окне i к состоянию k в следующем окне, при условии что оба этих состояния были вокализованными, выражается как

$$\Delta_{i,j,k} = G_1 \cdot \min \left[\xi_{j,k}, (G_2 + |\xi_{j,k} - \ln 2|) \right], \quad (9)$$

где

$$\xi_{j,k} = \left| \ln \frac{L_{i,j}}{L_{i-1,k}} \right|, \quad 1 \leq j < I_i; \quad 1 \leq k < I_{i-1}. \quad (10)$$

В том случае, когда текущее и последующее состояния являются невокализованными, имеем:

$$\Delta_{i,I_i,I_{i-1}} = 0. \quad (11)$$

Если же состояния для текущего и последующего окна отличаются, то в случае перехода от вокализованного к невокализованному получим

$$\Delta_{i,I_i,k} = H_1 + H_2 \cdot S_i + H_3 \cdot r_i, \quad 1 \leq k < I_{i-1}, \quad (12)$$

а для перехода от невокализованного к вокализованному —

$$\Delta_{i,j,I_{i-1}} = H_1 + H_2 \cdot S_i + H_3/r_i, \quad 1 \leq j < I_i. \quad (13)$$

Здесь

$$r_i = \frac{R(i,h)}{R(i-1,h)}, \quad (14)$$

$$R(i,h) = \sqrt{\frac{\sum_{j=0}^{J-1} (W_j S_{j+m+h})^2}{J}}, \quad m = iz, \quad (15)$$

где W — окно Хэмминга длиной $J = 00.3F_s$:

$$W_j = 0.54 - 0.46 \cos\left(\frac{2\pi j}{J}\right), \quad (16)$$

h — величина, контролирующая расстояние между центрами текущего и предыдущего окна для вычисления $R(i, h)$; S является функцией, обратной расстоянию Итакуры D_{IT} [13]:

$$S_i = \frac{0.2}{D_{IT}(i, i-1) - 0.8}. \quad (17)$$

При этом порядок линейного предсказания выбирается как

$$p = 2 + \lceil F_s / 1000 \rceil, \quad (18)$$

где $\lceil \cdot \rceil$ обозначает округление до ближайшего целого.

Обратимся теперь к анализу выражений для функций стоимости (12)—(13). Такой выбор стоимости переходов между состояниями обеспечивает её убывание в том случае, когда спектр сигнала подвержен быстрым изменениям, как в случае между границами вокализованных классов. Положительная константа H_1 является штрафом за изменение вокализованного состояния, вне зависимости от изменений в речевом сигнале. Это является отражением того факта, что смена вокализованных состояний в речевом сигнале происходит довольно редко.

Оптимальная траектория находится следующим образом:

$$D_{i,j} = d_{i,j} + \min_{k \in I_{i-1}} [D_{i-1,k} + \Delta_{i,j,k}], \quad 1 \leq j \leq I_i \quad (19)$$

с начальными условиями

$$D_{0,j} = 0, \quad 1 \leq j \leq I_0; \quad I_0 = 2. \quad (20)$$

Для каждого состояния в каждом окне необходимо хранить индексы для обратного прохода:

$$q_{i,j} = k_{\min}, \quad (21)$$

где k_{\min} — индексы k , минимизирующие $D_{i,j}$ в каждом окне.

Значение частоты основного тона в окне i определяется как

$$F_{0i} = \frac{F_s}{L_{i,j}}, \quad (22)$$

где значения j — это те величины, для которых достигается глобальный минимум D .

В проводимых нами расчётах использовались следующие значения параметров, фигурирующих в формулах (5)—(13):

$$F_s = 22050 \text{ Гц}, F_{ds} = 2205 \text{ Гц}, F_{0\max} = 450 \text{ Гц}, \delta_1 = 0.3, N_1 = 20, d_0 = 0,$$

$$\beta_0 = 0.3, G_1 = 0.2, G_2 = 0.35, H_1 = 0.005, H_2 = 0.5, H_3 = 0.5.$$

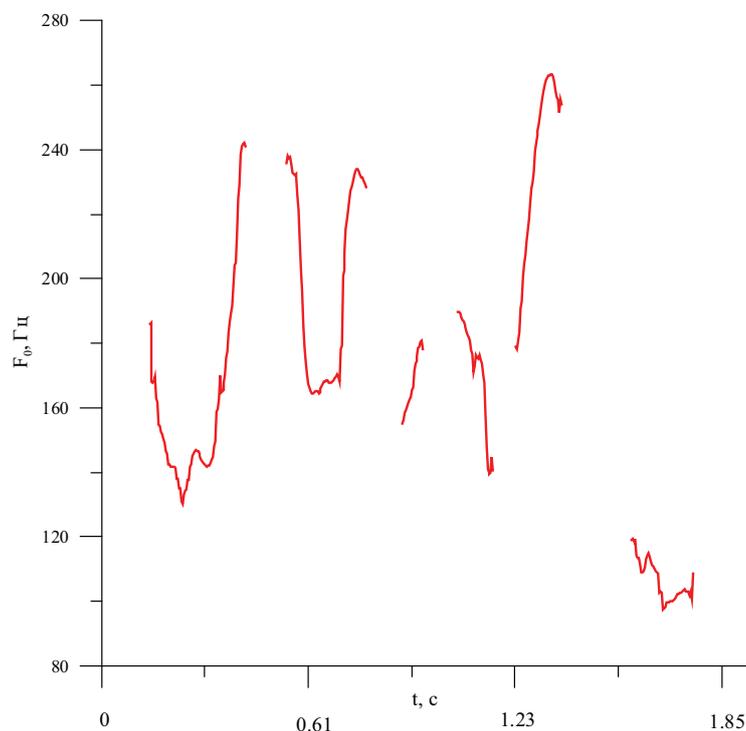


Рис. 1. Траектория частоты основного тона для отдельного фрагмента речевого сигнала

На рис. 1 представлены результаты применения описанного выше алгоритма для определения частоты основного тона для отдельного фрагмента речевого сигнала.

2.2. Определение значений первых трёх формант

Для определения необходимых нам первых трёх формант мы воспользуемся методом линейного предсказания [14]. В большинстве случаев, после того как коэффициенты линейного предсказания посчитаны, буфер коэффициентов дополняется единицей слева и нулями до степени числа 2 справа и подаётся на вход быстрому преобразованию Фурье (БПФ). Дополнение нулями нужно для повышения точности, так как при длине буфера меньше определённого значения (а в данном случае в буфере первоначально всего 20 коэффициентов) БПФ даёт очень неточные результаты. В результате применения быстрого преобразования Фурье и последующего преобразования составляющих спектра получается огибающая спектра сигнала, максимумы которой представляют собой формантные частоты.

Однако данный метод является весьма неточным и не позволяет надёжным образом определять значения формантных частот. Поэтому мы применяли другой способ нахождения формант, который основан на нахождении комплексных корней полинома [15]:

$$A(z) = 1 - \sum_{k=1}^p \alpha_k z^{-k} = \prod_{k=1}^p (1 - z_k z^{-1}), \quad (23)$$

где z — вектор коэффициентов линейного предсказания $[-1 \ a_1 \ a_2 \ \dots \ a_p]$.

Данные корни могут быть найдены с помощью метода Лагерра, имеющего кубическую сходимость для изолированных корней и линейную сходимость для кратных корней. Текущая итерация берётся в качестве корня, когда значение полинома при этой итерации меньше, чем вычисленная граница ошибок округления при нахождении значения полинома для этой итерации. Степень исходного полинома уменьшается, когда найден вещественный корень или пара комплексных корней, после чего итерационный процесс применяется к полиному уменьшенной степени [16].

Найденные корни многочлена могут быть представлены в следующем виде:

$$z_k = \exp([- \pi b_k + j2\pi F_k]/F_s), \quad (24)$$

где через b_k , F_k и F_s обозначены частотный диапазон, центральная частота k -й форманты и частота дискретизации соответственно. Так как α_k являются действительными, все комплексные корни имеют сопряжённые им, т.е. если (b_k, F_k) — корень, то $(b_k, -F_k)$ также представляет собой корень. Все b_k всегда положительны, так как для устойчивого предсказателя корни должны лежать внутри единичного круга ($|z_k| < 1$). Действительные корни не учитываются при нахождении формант, а комплексные корни упорядочиваются по возрастанию положительных F_k . Кроме того, исключаются корни, у которых частотная полоса превышает 200 Гц.

Указанный метод позволяет достаточно надёжно определять значения формант, однако в ряде случаев, как, например, при переходе от вокализованных к невокализованным участкам, а также при смене звучания фоном, возможно некорректное нахождение формант. Поэтому был использован метод динамического программирования, позволяющий учесть временные профили формант и тем самым значительно повысить точность их определения.

Введём базовые значения первых трёх формант F_{ni} , $i = 1, 2, 3$: $F_{n1} = 500$ Гц, $F_{n2} = 1500$ Гц, $F_{n3} = 2500$ Гц. Наша задача заключается в выборе среди N формант на протяжении K окон. В каждом окне k существует N_k способов отнести возможных кандидатов к определённым формантам:

$$N_k = \frac{n!}{(n-N)!N!}, \quad (25)$$

где n — число кандидатов формант в предыдущем окне, а N — рассматриваемое число формант.

Форманты выбираются из числа кандидатов по принципу минимальной стоимости, которая зависит от локальной стоимости, стоимости изменения частоты и стоимости перехода. Локальная стоимость λ_{kl} l -го назначения в k -ом окне зависит от частотного диапазона b_{kln} и отклонения от базового значения формант F_{nn} :



$$\lambda_{kl} = \sum_{n=1}^N \left[\beta_n b_{kln} + v_n \mu_n \frac{|F_{kln} - F_{ml}|}{F_{ml}} \right], \quad (26)$$

где β_n — стоимость увеличения частотного диапазона для n -й форманты, v_n — вероятность того, что данное окно является вокализованным, а μ_n определяет стоимость отклонения от базового значения n -й форманты.

Стоимость изменения частоты ξ_{kljn} между l -ым назначением в k -м окне и j -ым назначением в $(k-1)$ -м окне для n -й форманты выражается следующим образом:

$$\xi_{kljn} = \left[\frac{F_{kln} - F_{k-1jn}}{F_{kln} + F_{k-1jn}} \right]^2. \quad (27)$$

Стоимость перехода δ_{klj} определяется как

$$\delta_{klj} = \psi_k \sum_{n=1}^N \alpha_n \xi_{kljn}, \quad (28)$$

где α_n задаёт относительную стоимость изменения частоты между окнами для n -й форманты. Член ψ_k контролирует степень непрерывности траектории форманты:

$$\psi_k = \frac{\mathfrak{R}_k}{\max_{i \in K} \mathfrak{R}_i}, \quad (29)$$

где \mathfrak{R}_k — среднеквадратичное значение сигнала в k -м окне.

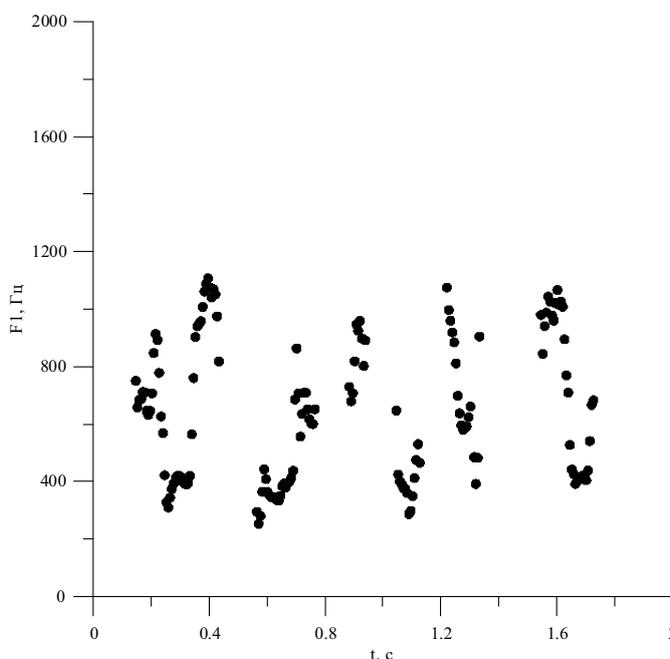


Рис. 2. Изменение первой форманты во времени для того же фрагмента речи, что и на Рис. 2.

Окончательно минимальная общая стоимость выбора кандидатов частоты форманты на протяжении K окон с N_k возможными соответствиями (назначениями) в каждом окне выражается следующим образом:

$$C = \sum_{k=1}^K \min_{l \in N_k} D_{kl}. \quad (30)$$

Здесь стоимость D_{kl} задаётся как

$$D_{kl} = \lambda_{kl} + \min_{j \in N_{k-1}} \kappa_{klj}, \quad (31)$$

где κ_{klj} — стоимость перехода от j -го назначения в $(k-1)$ -м окне к l -му назначению в k -м окне:

$$\kappa_{klj} = \delta_{klj} + D_{k-1j}. \quad (32)$$

На рис. 2 показаны результаты применения указанного алгоритма для вычисления первой форманты для отдельного фрагмента речевой волны. Из рисунка следует, что значения первой форманты лежат в пределах от 300 до 1200 Гц.

2.3. Признаки, основанные на вычислении кепстра

Существует немало частотных характеристик речевого сигнала, однако устойчивых к внешним шумовым помехам и позволяющих адекватно описывать данный отрезок сигнала среди них — единицы. Одними из наиболее успешно применяемых на практике являются кепстральные коэффициенты нелинейного масштаба [17]. Они определяются как действительный кепстр кратковременного сигнала, полученный из Фурье-преобразования данного сигнала. Отличие от действительного кепстра состоит в том, что используется нелинейная частотная шкала. В процессе построения программной системы нами были использованы первые 14 коэффициентов.

Амплитуда речевого сигнала существенно изменяется во времени. В частности, амплитуда невокализованных сегментов речевого сигнала значительно меньше амплитуды вокализованных сегментов. Подобные изменения амплитуды хорошо описываются с помощью функции кратковременной энергии сигнала. Кроме того, при изменении функционального состояния диктора происходит перераспределение энергии сигнала из одних частотных диапазонов в другие. Это приводит к тому, что среди признаков речевого сигнала, использующихся для выявления различных эмоциональных состояний, должны быть энергетические характеристики. Для получения энергии в определённых частотных диапазонах необходимо сначала вычислять свёртку исходного сигнала с полосовыми фильтрами, а затем уже подсчитывать энергию.

В качестве фильтра мы использовали так называемый Windows-Sinc-Blackman фильтр, имеющий следующую импульсную характеристику:

$$h(i) = K \frac{\sin(2\pi f_c(i - M/2))}{i - M/2} \left[0.42 - 0.5 \cos\left(\frac{2\pi i}{M}\right) + 0.08 \cos\left(\frac{4\pi i}{M}\right) \right], \quad i = 0, \dots, M$$



где K — константа нормировки, которая выбирается так, чтобы $\sum_{i=0}^M h(i) = 1$. Отметим, что M должно быть чётным числом.

3. Используемые характерные признаки

Для разделения всех речевых сигналов на два класса, соответствующих нормальному и стрессовому состояниям, с помощью методов, описанных в предыдущих разделах, были выделены 211 характерных признаков. В результате для каждого отдельного речевого сигнала получается вектор, размерность которого равна 211.

Для проведения испытаний качества работы разработанных методов и алгоритмов, реализованных в программной системе, необходимо наличие базы данных (БД), в которой хранятся речевые сигналы испытуемых, находящихся в различных эмоциональных состояниях. Проведённый анализ показал, что к настоящему моменту существует лишь одна такая русскоязычная база данных — RUSLANA (RUSsian LANguage Affective speech), созданная в университете Мейкай в Японии [18]. Она состоит из 3660 предложений, произнесённых 49 женщинами и 12 мужчинами в возрасте от 16 до 28 лет, для которых русский язык является родным. К сожалению, указанная БД недоступна для использования.

Среди англоязычных БД, безусловно, выделяется SUSAS (Speech Under Simulated and Actual Stress), собранная в течение нескольких лет в Колорадском университете в г. Боулдер (США) [19]. БД содержит записи 32 человек в возрасте от 22 до 76 лет, в том числе записи переговоров пилотов военных вертолётов, выполняющих боевые задания и, естественно, находящихся в стрессовых условиях. Отметим также БД из Северной Ирландии Belfast Natural Database [20] и из Массачусетского технологического института (США) [21].

Существует довольно большое число эмоциональных БД на немецком, японском и голландском языках [22—25]. В большинстве случаев в указанных БД рассматриваются следующие виды эмоций (приведены в порядке убывания частоты их использования): злость, печаль, радость, страх, отвращение, удивление, скука, стресс, презрение и др. Для свободного использования фактически доступны только две БД. Одна из них, имеющая очень небольшой размер, создана в Институте электроники и телекоммуникаций в г. Марибор (Словения) и содержит записи на четырёх языках — английском, французском, испанском и словенском [26]. Вторая БД — Berlin Database of Emotional Speech (EmoDB) — является наиболее подходящей для целей нашего исследования [27]. Она состоит из речевых фрагментов, произнесённых 5 мужчинами и 5 женщинами на немецком языке. Каждый из испытуемых произносил 10 фраз (например, «Скатерть лежит на холодильнике» или «Они только что отнесли это наверх и сейчас идут обратно»), пытаясь симитировать 6 различных эмоциональных состояний: злость, скука, отвращение, беспокойство/страх, радость, печаль. Кроме этого, дикторы произносили фразы нормальным спокойным голосом. Для некоторых эмоций в БД существует несколько вариантов их озвучивания одним и тем же лицом.

Тестирование разработанного программного комплекса осуществлялось на данной БД. Был создан специальный модуль, позволяющий считывать каждый звуковой файл из БД по отдельности и записывать выделенные характерные признаки в единый файл. После этого все вектора характерных признаков были разделены в соотношении 70/30. Первая часть была использована для обучения распознающей системы, работающей по методу опорных векторов. Оставшиеся вектора служили в качестве тестовой последовательности. Использовался метод 10-кратной перекрёстной валидации, при котором меняются последовательности, служащие в качестве обучающих и тестовых. Рассматривалось два класса: нормальное состояние и отклонение от нормы, куда попали все 6 имеющихся эмоциональных состояний. Особую трудность настройке распознающей системы придаёт тот факт, что в рассматриваемой БД эмоционально окрашенные файлы составляют порядка 90%. Таким образом, обучающая последовательность содержит значительно больше звуковых файлов, произнесённых в состоянии, отличном от нормального.

4. Классифицирующая система

Разрабатываемая система должна обладать возможностью относить речевой сигнал к одному из двух классов — в зависимости от того, был он произнесён в нормальном состоянии или отличном от него. Для построения подобного классификатора наилучшим образом подходит метод опорных векторов [28—29], позволяющий вычислить оптимальную разделяющую поверхность. Пусть нам известны N примеров $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, где $x_i \in R^n$ — векторы характерных признаков, а $y_i \in \{-1, +1\}$ — переменная, характеризующая эмоциональное состояние диктора: значение $y_i = -1$ соответствует нормальному состоянию, в то время как $y_i = +1$ означает состояние, отличное от нормального. В качестве разделяющей поверхности мы использовали гиперплоскость

$$(\mathbf{w} \cdot \mathbf{x}) - b = 0, \quad \mathbf{w} \in R^n, \quad b \in R, \quad (34)$$

поскольку она обладает наилучшими экстраполирующими свойствами по сравнению с нелинейными поверхностями. Соответствующий классификатор имеет следующий вид:

$$f(\mathbf{x}) = \text{sgn}((\mathbf{w} \cdot \mathbf{x}) - b), \quad (35)$$

т.е. в соответствии со знаком правой части вектор характерных признаков \mathbf{x} относится к тому или иному множеству.

Значения коэффициентов (\mathbf{w}, b) гиперплоскости находятся на основе обучающего множества в результате максимизации следующего функционала Лагранжа [28]:

$$L(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i, \mathbf{x}_j) \quad (36)$$

при дополнительных ограничениях

$$\alpha_i \geq 0, \quad \sum_{i=1}^N \alpha_i y_i = 0, \quad i = 1, \dots, N. \quad (37)$$

Здесь α_i — множители Лагранжа. Величина коэффициента b определяется из условий Куна-Таккера:



$$\alpha_i [y_i ((w \cdot x_i) - b) - 1] = 0, \quad i = 1, \dots, N. \quad (38)$$

Процесс нахождения коэффициентов (w, b) обычно называется настройкой или обучением системы.

В результате настройки точность распознавания состояния говорящего, т.е. отнесения к одному из двух возможных классов, составила 97.2%. Необходимо отметить, что это является одним из лучших на сегодняшний день результатов по эмоциональной классификации речевых сигналов. Типичным для большинства имеющихся двухклассовых систем распознавания является уровень порядка 70—80%.

5. Сокращение числа параметров

Для описанной системы существуют приложения, для которых очень важными характеристиками является время обработки речевого сигнала, а также объём используемой памяти: например, если система будет встроена в мобильный телефон, то это позволит во время разговора определять эмоциональное состояние собеседника. Для повышения производительности можно попытаться уменьшить число обрабатываемых параметров (разумеется, если это приведёт к незначительному снижению качества распознавания). Кажется естественным, что в первую очередь нужно отбросить характеристики сигнала, которые оказывают наименьшее влияние на результат. Как показали проведённые исследования [30], наилучшим образом значимость параметров характеризуется величиной его вариации, которая определяется следующим образом.

Приведём все компоненты векторов признаков к диапазону $[0, 1]$ и выберем k -ю компоненту: $k = 1, \dots, n$. Для каждого из заданных векторов признаков x_i , $i = 1, \dots, N$, вычислим расстояние d_i^k до найденной разделяющей гиперплоскости вдоль k -й компоненты:

$$d_i^k = |x_H^k - x_i^k|, \quad (40)$$

где

$$x_H^k = \frac{b - \sum_{j=k}^n w_j x_i^j}{w_k}, \quad (41)$$

а x_i^k — k -я компонента i -го вектора признаков x_i . Вариация v_k k -й компоненты определяется по формуле

$$v_k = \frac{1}{N} \sum_{i=1}^N d_i^k. \quad (42)$$

Чем меньше величина вариации, тем больше значимость соответствующего признака, поэтому степень значимости p_k можно выразить следующим образом (для удобства — в процентах):

$$P_k = \frac{v_k}{\sum_{k=1}^n v_k} \times 100\% . \quad (43)$$

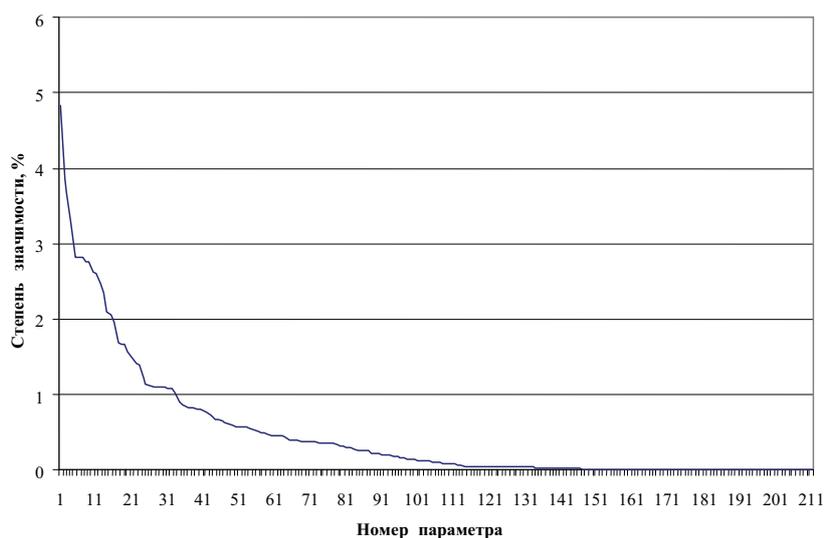


Рис. 3. Распределение степени значимости параметров

На рис. 3 приведено распределение степени значимости используемых 211 параметров речевого сигнала. Видно, что влияние на результат более чем половины параметров не превосходит 0,5%. Это обстоятельство позволяет сократить число параметров. Нами были оставлены параметры, значимость которых превосходит следующие пороговые значения: 0,5%, 1%, 1,5%, 2%. В следующей таблице (см. табл. 1) приведена зависимость точности определения эмоционального состояния дикторов от числа используемых параметров.

Таблица 1

Зависимость точности распознавания от числа параметров

Важность параметров , %	Число параметров, шт.	Точность распознавания, %
>0	211	97,2
>0.5	57	96,1
>1.0	34	95,9
>1.5	29	89,6
>2.0	15	86,7

Из таблицы видно, что сокращение числа параметров с 211 до 57 приводит к снижению точности лишь на 1,1%, а всего при 15 параметрах точность остаётся достаточно приемлемой для многих приложений. В таблице 2 перечислены наиболее важные для рассматриваемой задачи 15 параметров.



Таблица 2

Степень значимости наиболее важных параметров

Наименование параметра	Степень значимости, %
Медиана 12-го кепстрального коэффициента	4,84
Интерквартильный размах 7-го кепстрального коэффициента	3,86
Интерквартильный размах 6-го кепстрального коэффициента	3,68
Минимальное значение 10-го кепстрального коэффициента	3,27
Интерквартильный размах 5-го кепстрального коэффициента	2,82
Среднее значение 12-го кепстрального коэффициента	2,81
Медиана 7-го кепстрального коэффициента	2,81
Медиана 3-го кепстрального коэффициента	2,77
Интерквартильный размах 13-го кепстрального коэффициента	2,76
Минимальное значение 6-го кепстрального коэффициента	2,62
Минимальное значение 13-го кепстрального коэффициента	2,60
Среднее значение 1-го кепстрального коэффициента	2,47
Интерквартильный размах 1-го кепстрального коэффициента	2,34
Интерквартильный размах 4-го кепстрального коэффициента	2,09
Интерквартильный размах 7-го кепстрального коэффициента	2,05

Интересно, что все указанные параметры вычисляются на основе кепстра, что является весьма удачным обстоятельством с точки зрения повышения быстродействия системы. К сожалению, дальнейшее сокращение числа параметров приводит к резкому снижению достоверности классификации, что позволяет утверждать, что для рассматриваемого множества данных данное число параметров является минимально допустимым.

6. Заключение

Исследования, проведённые авторами при выполнении настоящей работы, показали высокую эффективность метода определения изменений в эмоциональном состоянии человека на основе анализа речевого сигнала. Достигнутая точность в 97.2% позволяет использовать такую систему для вынесения экспертных заключений, например, в бесконтактных «детекторах лжи», которые могут использоваться в финансовых учреждениях при выдаче кредитов. Сильно усечённый вариант системы, использующий на порядок меньшее число признаков, характеризующих речевой сигнал, требует незначительных вычислительных ресурсов, обеспечивая при этом точность 86.7%, чего вполне достаточно для интегрирования в бытовую технику, например, в мобильные телефоны.

Литература

1. L.Rothkrantz et al. Voice Stress Analysis. Text, Speech and Dialogues, ISBN 3-540-23049-1, *Lecture Notes in Artificial Intelligence*, P. 449—456, Springer, Berlin-Heidelberg-New York, 2004.
2. O-W.Kwon et al. Emotion Recognition by Speech Signals. In: *Proc. Intern. Conf. EUROSPEECH 2003*, Geneva. P. 125—128, 2003.
3. G.Zhou, J.H.L. Hansen, J.F. Kaiser. «Classification of Speech under Stress Based on Features Derived from the Nonlinear Teager Energy Operator», In: *Proc. IEEE Inter. Conf. on Acoustics, Speech, Signal Processing*, vol.I. P. 549—552, Seattle, 1998.
4. Zhou G.; Hansen J.H.L.; Kaiser J.F. «Methods for Stress Classification: Nonlinear TEO and Linear Speech Based Features». In: *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.4. P. 2087—2090, 1999.
5. B.D. Womack, J.H.L. Hansen. «Classification of Speech Under Stress using Target Driven Features», *Speech Communication, Special Issue on Speech Under Stress*, vol.20(1-2). P. 131—150, 1996.
6. G. Zhou, J.H.L. Hansen and J.F. Kaiser. «Nonlinear Feature Based Classification of Speech under Stress», *IEEE Transactions on Speech & Audio Processing*, 1997.
7. M.Sigmund. «Spectral Analysis of Speech under Stress». *Int. Journal of Computer Science and Network Security*, vol.7. P. 170—172, 2007.
8. W.J.Hess, *Pitch Determination of Speech Signals-Algorithms and Devices*, Springer-Verlag, Berlin, 1983.
9. L.R.Rabiner, M.J.Cheng, A.E.Rosenberg and A.McGonegal. «A comparative study of several pitch determination algorithms», *IEEE Trans. On Acoustics, Speech and Signal Processing*, Vol.ASSP-24: 399—413.
10. Y.Tadokoro, W.Matsumoto and M.Yamaguchi. «Pitch Detection of Musical Sounds Using Adaptive Comb Filters Controlled by Time Delay», In: *Proc. of the International Conf. on Multimedia and Expo*. P. 109—12, 2002.
11. A.Cheveigne and H.Kawahara. «YIN, a Fundamental Frequency Estimator for Speech and Music», *The Journal of the Acoustical Society of America*, vol. 111, Issue 4, pp.1917—30, 2002.
12. Беллман Р., Энджел Э. Динамическое программирование и уравнения в частных производных, М.: Мир, 1974, с. 208.
13. F.Zheng, Z.Song, L.Li, W.Yu, F.Zheng, W.Wu. The Distance Measure For Line Spectrum Pairs Applied to Speech Recognition, In: *Proc. 5th International Conference on Spoken Language Processing*, Sydney, №. 0171, 1998.
14. Маркел Дж.Д., Грей А.Х. Линейное предсказание речи. М.: Радио и связь, 1980. 248 с.
15. L.Rabiner, B.-H.Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1995. 507 p.
16. Уилкинсон Дж.Х. Алгебраическая проблема собственных значений, «Наука», М., 1970.
17. X.Huang, A.Acerio, H.W.Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall, 2001. 1008 p.
18. V.Makarova and V.A.Petrushin. «RUSLANA: A Database of Russian Emotional Utterances», In: *Proc. 2002 Int. Conf. Spoken Language Processing (ICSLP 2002)*, Colorado. P. 2041—2044, 2002.
19. Linguistic Data Consortium (LDC), <http://www ldc.upenn.edu/>.
20. E.Douglas-Cowie, R.Cowie and M.Schroeder. «A New Emotion Database: Considerations, Sources and Scope», In: *Proc. ISCA (ITWR) Workshop Speech and Emotion: A conceptual framework for research*, Belfast. P. 39—44, 2000.
21. R.Fernandez and R.W.Picard. «Modeling Drivers' Speech Under Stress», In: *Proc. ISCA Workshop (ITRW) on Speech and Emotion: A conceptual framework for research*, Belfast, 2002.
22. F.Schiel, S.Steiner, U.Turk. «The Smartkom Multimodal Corpus at BAS», In: *Proc. Lang. Resources and Evaluation*, Canary Islands, 2002.
23. B.Wendt and H.Scheich. «The Magdeburger Prosodie-Korpus», In: *Proc. Speech Prosody Conf. 2002*, Aix-en-Provence. P. 699—701, 2002.
24. Y.Niimi, M.L. Kasamatu, T.Nishimoto and M.Araki. «Synthesis of Emotional Speech Using Prosodically Balanced VCV Segments», In: *Proc. 4th ISCA tutorial and Workshop on research synthesis*, Scotland, 2001.
25. S.J.L. Mozziconacci and D.J. Hermes. «A study of intonation patterns in speech expressing emotion or attitude: production and perception», *IPO Annual Progress Report 32*, P. 154—160, IPO, Eindhoven, 1997.
26. Emotional Speech, <http://www.elektronika.uni-mb.si/eSpeech/speech.html>.



27. *F.Burkhardt, A.Paeschke, M.Rolfes, W.Sendlmeier and B.Weiss.* A Database of German Emotional Speech, In: *Proc. Intern. Conf. Interspeech*, Lissabon, 2005.
28. *V.N.Vapnik.* The Nature of Statistical Learning Theory (Statistics for Engineering and Information Science), 2nd edition. New York, Springer-Verlag, 1999. 304 p.
29. *Валник В.Н., Червоненкис А.Я.* Теория распознавания образов. М.: Наука, 1973. 416 с.
30. *A.A.Lukianitsa, F.M.Zhdanov and F.S.Zaitsev.* «Analysis of ITER Operation Mode Using the Support Vector Machine Technique for Plasma Discharge Classification», *Plasma Physics and Control Fusion*, v.50, №6. P. 14, 2008.

Шишкин Алексей Геннадиевич

Старший научный сотрудник факультета вычислительной математики и кибернетики МГУ им. М.В. Ломоносова, кандидат физико-математических наук.

Область научных интересов: математическое моделирование, распознавание образов, адаптивные методы обработки сигналов различного вида, обработка изображений, распознавание речи, применение современных информационных технологий в научных исследованиях.

Автор свыше 90 научных работ, в том числе одного учебника и одной монографии.

Лукьяница Андрей Александрович

Старший научный сотрудник факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова, кандидат физико-математических наук.

Область научных интересов: вычислительные методы, распознавание образов, нелинейная оптимизация, искусственные нейросети, генетические алгоритмы, скрытые модели Маркова, обработка изображений, распознавание речи.

Автор более 80 научных работ, включая один учебник и три монографии.

Перцептивно значимые элементы редуцированных словоформ



Риехакайнен Е. И.

В статье описываются результаты эксперимента по восприятию изолированных редуцированных словоформ, извлечённых из записей русской спонтанной речи. Предметом исследования является выявление тех элементов фонологического уровня, на которые опирается слушающий при распознавании редуцированных словоформ. На материале русского языка согласные оказываются более перцептивно значимыми элементами, чем гласные. Наибольшей перцептивной значимостью обладают: порядок следования согласных, начальный одиночный смычный согласный и ряд устойчивых для восприятия согласных в любой позиции.

The results of the experiment in which subjects were listening to isolated reduced word forms, extracted from spontaneous dialogues in Russian, show that consonants are more reliable sources of information for a listener than vowels in the process of Russian speech recognition. The most important elements are consonant order, single initial stop consonant and some perceptually strong consonants in any position. These results are proved by the recognition of the same word forms in context.

Теоретическая база исследования

В процессе восприятия речи, т.е. при необходимости «приписывания языковой структуры речевому сигналу» [1: 53], слушающий может пользоваться различными стратегиями. Исследования на материале изолированных фонем, слогов и слов часто предполагают пофонемное распознавание¹ звучащего сигнала (см., например: [2], [3]). В моделях пофонемного распознавания осуществляется последовательный анализ всех фонем словоформы.

Однако в процессе распознавания естественной звучащей речи использование только фонологической информации не является преобладающей стратегией. В условиях ограниченности во времени и фонетического несовершенства и/или неполноты сигнала² слушающий предпочитает стратегии, сочетающие «грубый, т.е. неполный фонологический

¹ В данной работе термины «восприятие» и «распознавание» считаются синонимами.

² «Чем большей спонтанностью обладает текст, тем больше в нём участков неполного типа» [4: 6].



анализ с одновременным лексико-семантическим и грамматическим» [5: 185].³ В большинстве современных моделей восприятия речи осуществление неполного фонологического анализа описывается через понятие активации. Предполагается, что некоторые элементы словоформы являются более перцептивно значимыми, чем другие. В процессе распознавания словоформы в словаре слушающего активируется всё множество единиц, которое соответствует последовательности её перцептивно значимых элементов. Впоследствии из этого множества активированных словоформ выбирается единственно возможная с учётом частотности, лексико-семантических и грамматических параметров.

При такой трактовке выделение перцептивно значимых элементов словоформы является необходимым этапом описания процесса восприятия речи.

Решение данной проблемы предполагает рассмотрение, по крайней мере, двух вопросов:

- какие части слова (начало, конец, ударный слог, консонантный «скелет» и т.п.) наиболее надёжно воспринимаются слушающими;
- существуют ли фонемы, которые распознаются лучше, чем остальные, и какие дифференциальные признаки фонем являются наиболее перцептивно устойчивыми.⁴

Первый вопрос получил достаточно подробное освещение в литературе по восприятию речи и решается по-разному в различных моделях. Так, в ранней версии модели когорты [8] первоначальная активация лексических единиц осуществляется исключительно по началам слов. При этом не существует единого мнения относительно того, каким должен быть размер начальной подцепочки: называются варианты от 1–2 до 3–5 фонем или первого слога.⁵ Если идентификация словоформы по начальным элементам оказывается невозможной, то необходимо продолжать анализ и привлекать следующие элементы, пока цепочка не станет уникальной.

Конец слова, напротив, считается ненадёжной для восприятия частью: в процессе речевосприятия слушающий не распознаёт конечный согласный слова гораздо чаще, чем начальный. Однако ослышки, вызванные неверной интерпретацией одного из звуков слова, связаны, как правило, именно с заменой начального, а не конечного согласного [9].

Тот факт, что ударный гласный практически никогда не редуцируется и произносится в слове наиболее отчётливо,⁶ позволяет ряду исследователей считать ударный гласный (или ударный слог) наиболее перцептивно значимым элементом [11]. Значимость ударного слога в целом при этом объясняется его большей продолжительностью и большей интенсивностью по сравнению с другими слогами в слове. Противники данной гипотезы

³ Обращение к пофонемному распознаванию оправдано только при необходимости восприятия новых, незнакомых, слов.

⁴ Термин «перцептивная устойчивость» заимствован из [6]. Встречающийся в более ранних работах (см., например, [7]) термин «перцептивная помехоустойчивость» решено было не использовать, поскольку в данной статье речь идёт о распознавании словоформ в условиях отсутствия каких-либо значимых помех (шума, акцента и т.п.).

⁵ См. ссылки на соответствующие работы в [1: 40].

⁶ См. обзор типичных случаев редукции гласных в разговорной речи в [10: 43–48].

указывают на то, что выводы о перцептивной значимости ударного гласного и начала слова делаются преимущественно на материале английского языка, в котором часто именно первый слог является ударным. При этом невозможно определить, какой именно признак — ударность гласного или его положение в начале слова — оказывается решающим [12: 4—47]. О большей значимости именно начального положения слога свидетельствуют результаты экспериментов, описанных в [13]: выполнение заданий на восприятие со словами, в которых ударным являлся не первый слог, занимало у испытуемых больше времени и требовало большего объёма памяти, чем работа со словами, в которых ударным был первый слог.

Проблема значимости ударного гласного связана с определением значимости гласных в целом для надёжности распознавания словоформы. На материале английского и немецкого языков было показано, что при ослышках ошибки в восприятии согласных происходят намного чаще, чем ошибки в восприятии гласных, а ударные гласные вообще практически не подвергаются изменениям. На основе данного наблюдения делается вывод о том, что по крайней мере в рассмотренных языках гласные в слове (особенно ударные) воспринимаются лучше и являются более надёжным источником информации, чем согласные [14].

Авторы монографии «Русская разговорная речь» отмечают, что в русской спонтанной речи слово узнаётся главным образом по согласным, гласные же служат преимущественно для передачи ритмики слова [10: 40-41]. Следовательно, в данной работе согласные признаются более значимыми элементами для восприятия речи, чем гласные. А исследования распознавания слов русского языка при наличии разного рода помех (шум, тугоухость, восприятие синтезированных звуков в различных условиях и т.п.) показали, что наиболее значимым фонетическим фактором для восприятия слов оказывается «ударная гласная», в большинстве экспериментальных условий значим также фактор «длина слова в слогах» [15].

В [4] описан эксперимент, в котором испытуемым предлагались для восприятия наиболее частотные редуцированные словоформы, извлечённые из записей русской речи. Была проанализирована роль 5 групп признаков: ударный гласный, ритмическая структура, начальная фонема словоформы, конечная фонема словоформы, количество согласных в словоформе. Результаты исследования показали, что количество согласных является самым существенным для опознания словоформы вне контекста: «чем большее число согласных выпадает из словоформы при её реализации, тем хуже она распознаётся» [4: 129]. При опущении одного из согласных количество правильно опознанных словоформ снижается на 37%. Кроме того, на надёжность распознавания влияют также сохранность ударного гласного и количества слогов: при искажении одного из этих двух признаков количество правильно опознанных словоформ снижается на 32%.

Современные модели распознавания звучащей речи, принимая во внимание вышеописанные наблюдения, в той или иной степени основываются на понятии повсеместной активации (*radical activation*). Предполагается, что входной речевой сигнал, соответствующий слову *dog*, может активировать и слово *bog* на основании совпадающих гласного и конечного согласного, несмотря на то что начальные согласные этих слов различны, при этом слову *dog* будет отдано предпочтение в процессе распознавания [16: 595]. В.Б. Касевич называет все те признаки словоформы, на которые слушающий может опираться в процессе восприятия, «просодиями» слова и относит к ним, в частности, ритмическую структуру слова (число слогов и распределение ударений и тонов), интонационный контур, тип распределения в слове согласных, все виды гармонии и сингармонизма [5: 201—202, 353—354].



Рассмотрение воспринимаемых фонем с точки зрения их дифференциальных признаков связано в первую очередь с созданием признаковых моделей восприятия речи (см. обзор в [1: 15—29]). Подобные модели не нашли широкого применения, однако необходимость анализа восприятия дифференциальных признаков до сих пор признаётся многими исследователями. Так, достаточно широко распространены исследования, направленные на выявление наиболее устойчивых дифференциальных признаков. Как правило, в подобных исследованиях рассматриваются прежде всего дифференциальные признаки согласных.

Анализ ослышек в английском языке показывает, что самым перцептивно устойчивым признаком является признак «шумность vs. сонорность»: при неверной интерпретации согласного в слове шумный в большинстве случаев заменяется на другой шумный, а сонант — на другой сонант, тогда как изменения по остальным признакам происходят намного чаще (например, смычный шумный переднеязычный /t/ в слове *great* может заменяться на /p/, /k/, /d/, /f/, /ç/) [9].

Существуют работы, посвящённые анализу восприятия дифференциальных признаков согласных фонем, являющихся начальными компонентами открытых слогов. Ряд исследований на материале языков, отличных от русского, свидетельствует о том, что «в составе стимулов типа CV наиболее устойчивыми, вероятно, являются признаки участия голоса и назальности, в наибольшей степени подверженным изменению оказывается признак места образования» [6: 29]. Эксперименты же по восприятию открытых слогов, являющихся элементами связного текста, на русском материале показали, что наиболее устойчивыми являются признаки глухости-звонкости, шумности-сонорности и мягкости-твёрдости, тогда как признаки способа образования и активного артикулирующего органа являются менее устойчивыми для восприятия. Кроме того, в результате этих экспериментов было обнаружено, что самые большие сложности для опознания представляют согласные в составе сложных консонантных кластеров [6: 32—155].

При наличии разного рода помех существенными факторами оказываются звонкость-глухость (и в слогах, и в словах), шумность-сонорность (в основном, для слогов), место образования (для слогов), способ образования (для слов и слогов в ряде экспериментальных условий). Твёрдость-мягкость оказалась существенной лишь в небольшом числе экспериментов [15].

В настоящей статье рассматриваются результаты экспериментального исследования восприятия редуцированных словоформ, извлечённых из записей спонтанной русской речи. Одной из задач являлось выделение перцептивно значимых элементов словоформ.

Материал эксперимента

Под редуцированными словоформами в данной статье понимаются формы слов, которые на сегментном уровне представлены меньшим количеством элементов, чем в полном варианте, предусмотренном нормами кодифицированного литературного языка. Подобные словоформы являются неотъемлемой

составляющей спонтанной речи. В 9 спонтанных бытовых диалогах на русском языке⁷ общей продолжительностью звучания 16 минут 33 секунды, которые были выбраны в качестве исходного материала для проведения экспериментов, встретилось 459 редуцированных словоформ (что составляет примерно 18% от общего числа словоформ экспериментальных текстов).⁸ Многочисленность редуцированных словоформ делает их распознавание важным элементом восприятия спонтанной речи в целом, а их использование в экспериментах по восприятию речи — перспективным для описания данного процесса и, в частности, для выявления перцептивно значимых элементов словоформ.

Поскольку основной задачей описываемого эксперимента было сопоставление надёжности распознавания редуцированных словоформ при изолированном и контекстном предъявлении, тестовая последовательность состояла из двух частей: в первой части испытуемым предъявлялись расположенные в произвольном порядке изолированные словоформы, во второй — они же в контексте.

Для проведения эксперимента были отобраны 24 редуцированные словоформы: 22 словоформы в женском произнесении (4 диктора) и 2 — в мужском (1 диктор).⁹

Контекст, в котором предъявлялись словоформы во второй части тестовой последовательности, мог быть как правым, так и левым — в зависимости от того, что позволял исходный текст диалогов.

Каждый стимул повторялся три раза: в первой части с интервалом в 0,5 секунды, а во второй — в 1 секунду. Межстимульный интервал составил для первой и второй частей соответственно 3,5 и 10 секунд. Порядок предъявления словоформ в первой и второй частях тестовой последовательности не совпадал. Всего в эксперименте было представлено 48 стимулов (по 24 в каждой части), общая продолжительность эксперимента составила 9 минут 20 секунд.

В таблице 1 представлен список всех словоформ-стимулов, их транскрипция, составленная на основании инструментального анализа сигналов, а также контекст, в котором они предъявлялись во второй части эксперимента.

Таблица 1

Материал эксперимента

Стимул	Данные инструментального анализа*	Контекст
Ходит	kóit	Она ходит к психологу
Наверно(е)	nám ⁱ	Наверно , тоже чтобы как-то

⁷ Участниками диалогов были 7 женщин и 1 мужчина в возрасте от 23 до 26 лет, для которых русский язык являлся родным. Параметры записи: 44100 Гц; 16 бит. Записи диалогов были предоставлены нашими французскими коллегами, которые проводили исследования русской и французской спонтанной речи.

⁸ Подробное описание редуцированных словоформ, встретившихся в данных спонтанных диалогах, представлено в: [17: 6—16].

⁹ Поскольку в качестве материала эксперимента выступал континуум спонтанной речи, при составлении тестовой последовательности из всех словоформ, представленных в исходных диалогах, были выбраны только те, при выделении которых из текста не возникало технических трудностей: не было искажений начального и конечного звуков слова под воздействием звуков соседних слов (например, озвончения конечного согласного под влиянием начального звонкого следующего слова), смеха, посторонних шумов.



кажется	kažs	мне кажется , таких людей
открыть	atkr't	не может открыть фотографии
представляешь себе	ptěšyt	и он об этом не знает, представляешь себе
нравятся	nrec	мне очень нравятся ситуации
какую-то	kógyt	продала какую-то картину
очень	eč	не очень хорошее
последний	pašyn'	сказала в последний раз
тебе	t'e	она тебе рассказывает
следующий	'sl'eš':	на следующий день
километров	kájna	двести километров может ехать
целый	cée	целый лист белой бумаги
по-моему	rom	он нам, по-моему , принёс
слушай / слышь	slyš	слушай , ты про него что-нибудь слышала?
понимаешь	рыывáš	понимаешь , она на каждом шагу останавливается
а она говорит	aongýt	а она говорит : «Да, но это не в первый раз!»
фильмов	'f'myf	из серии фильмов ужасов
достаточно	tstáčna	достаточно светло
потому что	ptyšt	потому что налоговая инспекция может
буквально	b'ál'n	буквально за час
появился	p'lc	появился какой-то дополнительный коридор
в пятницу	p'an'c	нет, это было не в пятницу , это было в среду
просят	p'os't	просят прислать фотографии

* Во втором столбце данные представлены с помощью фонематической транскрипции, созданной на основе фонетической транскрипции, предложенной Л.В. Щербой (см. [18: 272—273], [19: 38]); звуки, записанные как верхний индекс, обозначают короткие звуки или их призвуки (/ʰ/).

Анализ транскрипции показывает, что в рассматриваемых словоформах наряду с выпадением звуков встречается и другой тип редукции: некоторые звуки (преимущественно гласные) в составе словоформ подвергаются количественным и качественным изменениям.

Далее в статье будут подробно рассмотрены результаты восприятия изолированных редуцированных словоформ, т.е. стимулов, вошедших в первую часть тестовой последовательности. В конце статьи будут приведены

некоторые результаты распознавания исследуемых словоформ в контексте (во второй части тестовой последовательности и в другом эксперименте), позволяющие судить о перцептивной значимости элементов редуцированных словоформ.

Методика эксперимента и испытуемые

Тестовая последовательность, записанная на магнитную ленту, предъявлялась испытуемым через громкоговорители.

Участникам эксперимента предлагалось прослушать предъявляемые стимулы и записать то, что они услышали, буквами русского алфавита. Разрешалось пропускать сигналы (ставить прочерк в анкете), если испытуемый не мог определить, что он слышит, или не успевал идентифицировать стимул.

В инструкции было указано только то, что испытуемым будут предъявлены фрагменты русской речи, поэтому испытуемые могли интерпретировать услышанное как любую последовательность звуков русской речи (необязательно осмысленную) и записывать в ответах не только слова и словосочетания, но и отдельные слоги и даже звуки.

Испытуемыми были студенты и преподаватели филологического и восточного факультетов Санкт-Петербургского государственного университета.

В эксперименте приняли участие 40 человек.

Результаты эксперимента

Восприятие изолированных словоформ

В ходе обработки полученных данных были подсчитаны процентные соотношения по следующим параметрам: количество распознанных словоформ, количество отказов, число осмысленных буквосочетаний в ответах испытуемых. Считалось, что словоформа распознана верно, если испытуемый записывал в ответе то слово, редуцированный вариант которого был представлен в эксперименте (например, тебе для /t'e/ или очень для /eč/). Кроме того, поскольку в эксперименте проверялась надёжность восприятия редуцированных словоформ с искажениями в основе, то при обработке полученных данных считалось, что словоформа распознана верно, даже если испытуемый записывал в бланк ответа флексию, отличную от исходной. Таким образом, в данном эксперименте правильными ответами считались: *нравится* (для стимула *нравятся*), *ходят* (для *ходит*), *просит* (для *просят*), *пятница* (для *(в) пятницу*). Для стимула *слушай* /slyš/ верным признавался и ответ *слышь*, поскольку редуцированные варианты этих вводных слов очень близки между собой и невозможно определить, какое же слово было произнесено на самом деле в диалоге, на материале которого составлялась тестовая последовательность.

При анализе перцептивной значимости элементов словоформ подсчитывался процент верных отражений звуков на соответствующих позициях в ответах испытуемых. В данном случае верными считались те ответы, в которых было записано то, что соответствовало транскрипции, представленной в таблице 1, а не то, что должно было звучать в редуцированном варианте соответствующей словоформы (например, *к*, а не *х* в начале



словоформы *ходит /kóit/* или *a*, а не *e* на месте ударного гласного в *наверно /námí/*). Для сопоставления результатов вычислялись доверительные интервалы на 5%-ном уровне значимости. Статистически были выделены 2 параметра: надёжность распознавания того или иного элемента словоформы и наличие преобладающего варианта при отражении этого элемента испытуемыми. Распознавание считалось надёжным, если испытуемые воспринимали реально звучащую фонему достоверно чаще, чем все остальные фонемы: проводилось сравнение доверительных интервалов для верного варианта и суммы всех неверных. Вариант считался преобладающим, если процент испытуемых, зафиксировавших его в своих ответах, был достоверно выше любого другого варианта для данного стимула: доверительный интервал для наиболее частотного варианта сопоставлялся с каждым из менее частотных вариантов для данного стимула.

Прежде всего, необходимо отметить, что процент распознавания редуцированных словоформ при изолированном предъявлении оказался очень низким. Так, 58,3% всех предъявляемых словоформ не были правильно опознаны ни одним испытуемым. В таблице 2 представлены результаты распознавания словоформ, которые были опознаны хотя бы одним испытуемым.

Таблица 2

Надёжность распознавания изолированных редуцированных словоформ

Стимул	Количество верных распознаваний с учётом доверительных интервалов, %
достаточно	4,1<12,5<26,8
просят	2,7<10,0<23,7
слушай	1,5<7,5<20,4
понимаешь	1,5<7,5<20,4
в пятницу	1,5<7,5<20,4
нравятся	0,5<5,0<17,0
очень	0,2<2,5<13,2
тебе	0,2<2,5<13,2
а она говорит	0,2<2,5<13,2
буквально	0,2<2,5<13,2

Только для двух словоформ распознавание превышает или равно 10%. Четыре словоформы были правильно опознаны только 1 испытуемым из 40. Таким образом, при анализе восприятия редуцированных словоформ, предъявляемых изолированно, нельзя говорить о надёжном распознавании ни одной из них.

Стимулы, вошедшие в тестовую последовательность, не были специально отобраны для проведения эксперимента по изучению перцептивной значимости элементов словоформ: по таблице 1 видно, что фонетический облик исследуемых словоформ достаточно разнообразен. Тем не менее, полученный в ходе эксперимента материал позволяет сделать ряд предвари-

тельных выводов о перцептивной значимости элементов редуцированных словоформ и устойчивости некоторых признаков фонем.

Опираясь на описанные выше гипотезы о перцептивной значимости различных элементов словоформы, рассмотрим, насколько надёжно были восприняты испытуемыми ритмическая структура и консонантный «скелет» (в частности, начальный согласный). В инструкции, предъявлявшейся испытуемым, не требовалось расставлять ударения, поэтому при оценке перцептивной значимости ритмической структуры будет проанализирована только надёжность распознавания количества слогов и того гласного, который является ударным; позиция же ударения не может быть оценена по данным проведённого эксперимента.

Инструментальный и слуховой анализ показали, что количество слогов практически во всех исследуемых редуцированных словоформах меньше, чем в их нередуцированных вариантах: реально звучат как трёхсложные только 3 стимула (*последний* /pasýni/, *понимаешь* /ryiváš/, *а она говорит* /aongýt/), как двусложные — 9 словоформ, большинство же словоформ (12) являются односложными. Причём в двусложных и трёхсложных словоформах безударные гласные часто являются очень короткими (иногда это характерно и для ударных гласных).¹⁰

Только для 13 стимулов (54,2%) можно выделить преобладающий вариант при отражении количества слогов в услышанной словоформе. Интересно отметить, что в стимуле *открыть* /atkrʹ/ большинство испытуемых (10 из 14 давших ответ) вообще не услышали гласного звука. Результаты для 13 стимулов представлены в таблице 3.

Таблица 3

Отражение количества слогов в ответах испытуемых

Стимул	Отражение кол-ва слогов в ответах испытуемых, чел.					Преобладающий вариант, %
	0 слогов	1 слог	2 слога	3 слога	4 слога	
ходит /kóit/	0	34	2	0	0	87,2
наверно /nám/	0	27	6	1	0	79,4
кажется /kažs/	4	33	0	0	0	89,2
нравятся /nrec/	0	34	2	2	0	89,5
какую-то /kógyt/	0	23	10	0	0	63,8
следующий /sl'eš':/		30	3			83,3
цельй /cée/	5	25	4	0	0	73,5
по-моему /pom/	1	35	2	0	0	92,1
слушай /slyš/	6	29	2	1	0	76,3
а она говорит /aongýt/	0	4	21	8	1	61,8
фильмов /f'ímýf/	0	31	4	0	0	88,6
достаточно /tstáčna/	0	3	24	4	6	63,2
появился /p'ic/	2	32	1	0	0	91,4

¹⁰ Столь значительная редукция ритмической структуры исследуемых словоформ, по-видимому, и является причиной их плохой распознаваемости при изолированном предъявлении (ср.: [4: 131—132]).



Серым цветом в таблице выделено число испытуемых для каждого из стимулов, которые верно отразили количество слогов; жирным шрифтом выделен тот вариант, который преобладал в ответах испытуемых. В последнем столбце отмечен процент, который составляет преобладающий вариант от общего числа полученных ответов.

В 7 из 13 стимулов большинство испытуемых услышали столько слогов, сколько в них звучало на самом деле, в шести — меньше, чем в предъявлявшемся стимуле. Для 6 из 7 стимулов, количество слогов, в которых было верно отражено испытуемыми, распознавание можно считать статистически надёжным. Все эти стимулы представляют собой односложные словоформы (/kəʒs/, /nrec/, /s'eš:/, /pom/, /slyš/, /p'lc/). Таким образом, только в 25% из всех предъявлявшихся в эксперименте стимулов количество слогов было отражено испытуемыми верно. Ни для одной из двух- и трёхсложных словоформ распознавание количества слогов не является надёжным.

Анализируя ударные гласные в предъявлявшихся стимулах, можно заметить, что редукция затрагивает не только безударные гласные, но и ударный. Данные инструментального и слухового анализа свидетельствуют о том, что ударный гласный сохраняется только в 15 словах тестовой последовательности (см. табл. 1). Следовательно, в ходе оценки перцептивной значимости ударного гласного будет рассмотрена надёжность распознавания того гласного, который находился в предъявляемом стимуле на позиции ударного, а не того, который должен быть на данной позиции при полном произнесении.

Гласный в ударной позиции был надёжно распознан только в 3 стимулах: *ходит* /kóit/, *просят* /p'os't/ и *по-моему* /pom/¹¹ (79,5%, 84,2% и 84,2% правильных опознаваний ударного гласного соответственно). Ещё для 5 стимулов можно выделить преобладающие варианты ответов испытуемых. Для словоформы *понимаешь* /p'yiváš/ преобладающий вариант (69,7% ответов) совпадает с той фонемой, которая звучала в предъявляемом стимуле (/a/). Для стимулов *слушай* /slyš/, *появился* /p'lc/ и *фильмов* /'f'ityf/ в ответах испытуемых преобладает фонема /u/ на месте ударного гласного (55,2%, 60% и 68,6% ответов соответственно). В стимуле *а она говорит* /aongýt/ преобладающим вариантом интерпретации ударного гласного является фонема /e/ (38,2%).

Можно предположить, что перцептивно важным должен быть единственный гласный в односложном стимуле, поскольку гласный традиционно считается вершиной слога. Однако, кроме уже описанных выше результатов по стимулам *просит* и *по-моему*, вряд ли можно говорить об однозначном распознавании гласных в односложных стимулах.

Таким образом, ни количество слогов, ни ударный гласный не являются на проанализированном материале перцептивно значимыми признаками словоформы: испытуемые склонны распознавать в предъявляемом стимуле

¹¹ В стимуле *по-моему*, который звучит в эксперименте как /pom/, не ясно, является гласный /o/ исходным ударным или безударным гласным первого слога, но для изучения надёжности восприятия гласных этот вопрос не является столь принципиальным.

меньше гласных, чем в нём содержится на самом деле, а ударный гласный был надёжно распознан только в 3 стимулах.

В рамках описываемого в настоящей статье исследования не проводилось последовательного анализа перцептивной значимости всех признаков гласных, однако необходимо отметить, что, по-видимому, особенную важность в процессе речевосприятия приобретает признак огубленности. Этот вывод подкрепляется следующими наблюдениями: — хотя в целом ударный гласный воспринимается испытуемыми не очень надёжно, в тех словах, где он является огубленным (/o/ или /u/), распознавание ударного гласного выше, чем в большинстве других слов (при этом иногда испытуемые могут воспринимать /o/ вместо /u/ и наоборот, но признак огубленности сохраняется); — в ряде слов большинство испытуемых восприняло ударный гласный как огубленный под влиянием соседних губных согласных: например, *фильмов* /fⁱmyf/ (гласный был интерпретирован как **у** или **ю** 54,3% испытуемых: например, *чуф*, *тюф* и др.), *буквально* /b^uál'n/ (буквы **у** или **о** на позиции гласного встретились в ответах 75,7% испытуемых: *бун*, *больн* и др.), *появился* /pⁱl's/ (**у** или **о** на месте единственного гласного в словоформе — 62,9%: *пуск*, *возг* и др.).

Данные наблюдения являются, несомненно, интересными, но при их интерпретации необходимо учитывать, что огубленный гласный может появляться в некоторых случаях по законам фонотактики.

В целом, сделанные выше наблюдения свидетельствуют, по-видимому, о том, что на материале русского языка гласные не являются наиболее перцептивно значимыми элементами слова (по крайней мере, при восприятии редуцированных словоформ).

При анализе восприятия консонантного «скелета» необходимо рассмотреть несколько аспектов. Во-первых, стоит отметить, что, в целом, испытуемые достаточно точно отражают порядок следования согласных в исследуемых словоформах. Этот факт приобретает особую значимость, если принять во внимание то обстоятельство, что в ходе эксперимента словоформы предъявлялись триадами с небольшими временными интервалами внутри триады, что увеличивало вероятность неверного осмысления границ словоформы, а также влияния фонотактики.

Экспериментальный материал позволяет сделать некоторые выводы относительно того, насколько адекватно воспринимается начальный согласный в редуцированных словоформах. Для этого были рассмотрены все стимулы из тестовой последовательности, начинающиеся с одиночного согласного звука, а также с последовательности одиночного согласного и призвука другого согласного (всего 14 словоформ). Нужно отметить, что в 13 из анализируемых стимулов начальный согласный был смычным, в одном — аффрикатой (/сéе/). В шести стимулах данный согласный был распознан надёжно (выделен серым цветом в таблице 4), ещё для трёх верный вариант является преобладающим вариантом ответа.

Таблица 4

Словоформы, начальный согласный в которых был распознан хорошо

Словоформа	% верных распознаваний начального согласного	Примечания
ходит /kóit/	97,4	
какую-то /kóгыt/	91,7	



тебе /t'e/	70,3	Для данных словоформ получено примерно равное количество ответов для твёрдого и мягкого начального согласных. Оба варианта считались в данном случае верным распознаванием
километров /kájna/	88,6	
по-моему /pom/	60,5	
понимаешь /pъiváš/	60,6	
буквально /b'ál'n/	56,8	
(в) пятницу /p'an'c/	84,6	
просят /p'os't/	81,6	

Кроме того, стоит отметить надёжность распознавания согласного /g/ в стимуле *а она говорит* /aongýt/ (94,1%).

Плохо распознавался начальный согласный в словоформах *наверно* /námí/, *кажется* /kažs/, *последний* /pasýni/ и *появился* /p'ic/, а также аффриката /c/ в словоформе *целый* /céé/, на месте которой большинство испытуемых (52,9%) услышали щелевой /f/.

Сочетания согласных в начале слова распознавались плохо, что соотносится с данными о плохом распознавании консонантных кластеров, полученными в ходе экспериментов других исследователей [6]. Начальные консонантные кластеры были восприняты большинством испытуемых как одиночные согласные (часто щелевые).

Отдельно необходимо рассмотреть результаты распознавания начального согласного /n/ в стимулах *наверно* (звучал как /námí/) и *нравятся* /nrec/. В этих стимулах, вопреки распространённому мнению о том, что признак шумности-сонорности является одним из наиболее устойчивых перцептивных признаков, начальный сонорный /n/ воспринимается большинством испытуемых как шумный смычный /p/. И если в словоформе *нравятся* этот факт можно объяснить более высокой частотностью начального сочетания /pr/ по сравнению с сочетанием /nr/ (т.е. испытуемые, даже не зная, что им предлагаются для прослушивания слова, подсознательно ориентируются на законы фонотактики и правила построения слов русского языка), то для стимула *наверно* это объяснение не подходит.

Анализ перцептивно устойчивых признаков начального согласного на материале остальных стимулов подтверждает тезис о наибольшей устойчивости признаков шумности-сонорности и глухости-звонкости [6], хотя в ответах испытуемых встречались практически все возможные варианты замен согласных. Признак мягкости-твёрдости начальных согласных является менее надёжным для восприятия (в словоформах *тебе* и *километров* было получено примерно равное количество ответов с мягким и твёрдым начальным согласным), что согласуется с результатами по восприятию слогов и слов при наличии разного рода помех [15].

В позициях, отличных от начальной, замены согласных по различным признакам встречались чаще, чем в начале слова. Относительно этих позиций можно делать выводы скорее об устойчивости восприятия отдельных согласных,

чем об устойчивости определённых позиций в слове, поскольку консонантные «скелеты» исследуемых словоформ достаточно разнообразны. Наиболее последовательно воспроизводятся смычные шумные согласные (особенно в конце словоформы), щелевые /š/, /š':/, аффриката /č/, а также сонант /n/, являющийся суффиксом наречий (в словоформе *буквально* /b'ál'n/ в 86,5%, в словоформе *достаточно* /tstáčna/ — в 100% ответов). Хуже всего, как и в начале слова, распознаются консонантные кластеры, чаще всего они заменяются щелевыми согласными.

В качестве особо интересных случаев стоит отметить появление согласного /r/ в ответах на те стимулы, в составе которых он на самом деле отсутствует (данный согласный не представлен ни в нередуцированном, ни в редуцированном варианте соответствующих словоформ): *понимаешь* /ryiváš/ (63,6%, например, ответы *пэраш*, *пурваф*, *дурак*), *кажется* /kažs/ (51,4%: *кырс*, *кэрс*, *трэст* и др.), *целый* /céel/ (35,3%: *фёр*, *фер* и др.), *какую-то* /kógyit/ (30,6%: *курт*, *кург* и др.), *потому что* /ptyšt/ (27,6%: *трст*, *фирст* и др.), *(в) пятницу* /p'an'c/ (25,6%: *пер*, *перед* и др.), — а также достаточно последовательное отражение этого звука при восприятии тех стимулов, в которых он изначально присутствует.

Появление на месте ударного гласного в стимуле *кажется* /kažs/ гласного *о* в окружении *к_сть* или *г_рст* может также объясняться тем, что слушающий даже при работе с изолированными редуцированными словоформами, которые он не может соотнести с каким-либо конкретным словом, опирается на знание законов фонотактики и частотности звуко сочетаний в русском языке. Такая трактовка является дополнительным подтверждением того, что при восприятии речи слушающий опирается на консонантный «скелет» словоформ.

Таким образом, результаты проведённого эксперимента позволяют предположить, что при необходимости идентификации редуцированной словоформы в процессе восприятия речи слушающий опирается в первую очередь на консонантный «скелет» словоформы (последовательность согласных, начальный смычный согласный, некоторые устойчивые для восприятия согласные).

Восприятие редуцированных словоформ в контексте. Некоторые наблюдения

Выдвинутая гипотеза подтверждается результатами восприятия описанных словоформ в контексте. Прежде всего, уже при появлении небольшого контекста редуцированные словоформы распознаются испытуемыми достаточно надёжно (14 словоформ были распознаны более чем половиной испытуемых).

Сопоставление результатов распознавания словоформ при изолированном предъявлении и в контексте представлено на диаграмме 1.

При необходимости интерпретации стимула в контексте испытуемые достаточно легко восстанавливают количество слогов до того, которое должно быть в нередуцированном варианте. Изменение же согласных для придания осмысленности фразе встречается намного реже.

Отдельного внимания заслуживают результаты другого эксперимента, направленного на определение роли контекста в процессе восприятия. В этом эксперименте описанные выше редуцированные словоформы были помещены в контекст, который должен был спровоцировать их неверную интерпретацию. Например, словоформа *просит* /p'os't/

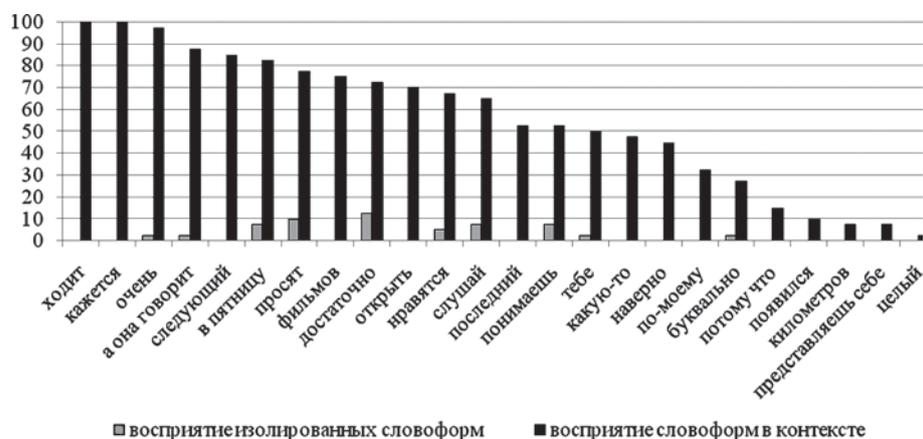


Диаграмма 1. Количество верных распознаваний редуцированных словоформ при изолированном предъявлении и в контексте (в процентах от общего числа испытуемых)

помещалась в контекст **просят раздувает** и **просят живёт** (ожидалось появление словоформы *пусть* в ответах испытуемых), словоформа *говорит /gɨt/* помещалась в контексты **говорит свечку**, **свечка говорит** (ожидалось появление ответов *горит*, *гасит* и т.п.) и т.д. Методика проведения данного эксперимента совпадала с методикой эксперимента по распознаванию редуцированных словоформ в естественном контексте (см. выше).

Результаты восприятия большинства стимулов в данном эксперименте подтверждают важность начального согласного как перцептивно значимого элемента словоформы, а также значимость консонантного «скелета». Так, словоформа *говорит /gɨt/* в контексте **свечка говорит** была воспринята 33,3% испытуемых как *горит*, для словоформы *просит /pʲosʲtʲ/* в обоих контекстах самым частотным вариантом ответа был *пусть*, при интерпретации словоформы *фильмов /ʲfʲimʲɨfʲ/* в контекстах и **фильмов машина** и и **фильмов приснился** испытуемые предпочитали вариант *Ефим(ов)(а)у* (12,8% и 23,1% соответственно), словоформа *тебе /tʲe/* в контексте **тебе рядом с этим** была воспринята 36% испытуемых как *ты*, 5,1% испытуемых как *то* и 3% испытуемых как *те* (правильный ответ *тебе* встретился у 6,1% испытуемых). Кроме того, интересен пример со словоформой *наверное /návʲnʲe/* в контекстах **то же самое делается**, **наверно** и **порошки лучше покупаются, наверно**, где начальный согласный в контексте в большинстве случаев воспринимается как /n/ (тогда как при изолированном предъявлении наиболее частотным был согласный /p/), однако из-за влияния фонотактики в ответах испытуемых появляются также слова *сами* и *сам*.

На количество слогов и качество гласных (в том числе ударного) испытуемые опираются, по-видимому, меньше, чем на согласные; в процессе интерпретации редуцированной словоформы они склонны заменять гласные и увеличивать количество слогов (например: замены гласных в словоформе *тебе /tʲe/* (/e/->/ы/), *просят /pʲosʲtʲ/* (/o/->/u/), «восстановление» слога с ударным гласным в словоформе *говорит /gɨtʲ/* в контексте **говорит свечку** (*гáсит*, *гláдит*), «восстановление» гласных в окончаниях: *фильмов /ʲfʲimʲɨfʲ/* (*Ефимову*, *Ефимова*), *следующий /sʲlʲeʃʲː/*).

Выводы

Результаты описанных выше экспериментов показывают, что редуцированные словоформы распознаются в естественном (неизменённом) контексте намного лучше, чем при изолированном предъявлении.

Количество слогов и ударный гласный отображаются испытуемыми менее надёжно, чем консонантный «скелет» словоформы, что, по-видимому, свидетельствует о том, что при распознавании редуцированных словоформ согласные являются более перцептивно значимыми элементами словоформы, чем гласные. Наибольшей перцептивной значимостью обладают: порядок следования согласных, начальный одиночный смычный согласный, смычные шумные согласные, щелевые /š/, /š':/, аффриката /č/ в любой позиции.

Наиболее перцептивно устойчивыми признаками начальных согласных являются шумность-сонорность и глухость-звонкость. Кроме того, результаты эксперимента указывают на перцептивную значимость признака огубленности гласных.

Таким образом, результаты, полученные в описанном исследовании, расходятся с результатами, полученными при анализе ослышек на материале английского и немецкого языков, которые указывают на перцептивную значимость гласных в слове.

В то же время данные эксперимента не указывают на то, что начальный согласный является наиболее перцептивно значимым элементом словоформы: существует ряд согласных, которые являются перцептивно значимыми вне зависимости от позиции; с другой стороны, в начале слова надёжно распознавались только одиночные смычные согласные, консонантные же кластеры воспринимались плохо как в начале словоформы, так и в любой другой позиции.

Следовательно, наиболее правдоподобным объяснением полученных результатов является модель повсеместной активации с опорой на консонантный «скелет».

Полученный материал позволяет предположить, что при необходимости распознавания редуцированной словоформы в спонтанной речи слушающий выбирает из перцептивного словаря ту словоформу, которая подходит по консонантному «скелету» и по контексту.

Безусловно, данная стратегия не является единственно возможной при восприятии речи. По всей видимости, в распоряжении слушающего находится целый набор стратегий, из которых он может выбрать наиболее рациональную в той или иной ситуации. Возможно также, что данный набор стратегий обусловлен конкретным языком и именно этим объясняются различия, полученные на материале английского и немецкого языков, с одной стороны, и русского языка — с другой. Представленная же в статье стратегия с опорой на консонантный «скелет» позволяет, по крайней мере, описать механизм распознавания редуцированных словоформ в русской спонтанной речи. Возможность применения данной стратегии в других ситуациях является предметом дальнейших исследований.

Литература

1. Венцов А.В., Касевич В.Б. Проблемы восприятия речи. М.: Едиториал УРСС, 2003.
2. Чистович Л.А., Кожевников В.А., Алякринский В.В. и др. Речь. Артикуляция и восприятие. М., Л.: Наука, 1965.



3. *Чистович Л.А., Венцов А.В., Гранстрем М.П. и др.* Физиология речи. Восприятие речи человеком. Л.: Наука, 1976.
4. *Бондарко Л.В., Вербицкая Л.А., Гейльман Н.И. и др.* Фонетика спонтанной речи / Под ред. Н.Д. Светозаровой. Л.: Изд-во Ленинградского университета, 1988.
5. *Касевич В.Б.* Труды по языкознанию: В 2-х т. Т.1 / Под ред. Ю.А. Клейнера. СПб: Филологический факультет СПбГУ, 2006.
6. *Ягунова Е.В.* Восприятие согласных фонем и их дифференциальных признаков (экспериментально-фонетическое исследование на материале русского языка) / Диссертация на соискание учёной степени кандидата филологических наук. Рукопись. СПб, 1994. 261 с.
7. *Штерн А.С.* Влияние лингвистических факторов на восприятие речи. / Диссертация на соискание ученой степени кандидата филологических наук. Рукопись. Л., 1981. 251 с.
8. *Marslen-Wilson, W.D., Welsh, A.* Processing interactions during word recognition in continuous speech // *Cognitive Psychology*. 1978. Vol. 1. N1. P. 29-63.
9. *Bond Z.S.* Slips of the Ear // *The Handbook of Speech Perception* / Ed. by D.B. Pisoni, R.E. Remez. Oxford: Blackwell, 2005. P. 290-310.
10. *Русская разговорная речь* / Под ред. Е.А. Земской. М.: Наука, 1973.
11. *Cutler A., Norris D.* The Role of Strong Syllables in Segmentation for Lexical Access // *Journal of Experimental Psychology: Human perception and performance*. 1988. Vol. 14, N1. P. 113-121.
12. *Protopapas A.* Perspectives of Syllables, Stress, and Interactions in Speech Perception: Experimental and Connectionist Approaches / Theses for the Degree of Doctor of Philosophy. Brown University, 1997. 183 p.
13. *Mattys S., Samuel A.* Implications of stress pattern differences in spoken word recognition // *Journal of Memory & Language*. Vol. 42, №4, 2000. P. 571—596.
14. *Aitchison J.* Words in the Mind: An Introduction to the Mental lexicon. Oxford: Blackwell, 2003.
15. *Штерн А.С.* Перцептивный аспект речевой деятельности: (Экспериментальное исследование). СПб: Издательство Санкт-Петербургского университета, 1992.
16. *Luce P.A., McLennan C.T.* Spoken Word Recognition: The Challenge of Variation // *The Handbook of Speech Perception*. / Ed. by D.B. Pisoni, R.E. Remez. Blackwell Publishing Ltd, Berlin, Oxford, 2005. P. 592—609.
17. *Горбова Е.В., Слепокурова Н.А., Комовкина Е.П., Макарова А.Б., Малов Е.М., Риехакайнен Е.И.* Современная русская спонтанная телевизионная речь: некоторые итоги исследования // *Современная русская речь: состояние и функционирование: Сборник аналитических материалов* / Под ред. О.И. Глазуновой, Л.В. Московкина, Е.Е. Юркова. СПб: Издательский дом «МИРС», 2008. С. 5—82.
18. *Бондарко Л.В.* Фонетика современного русского языка: Учебное пособие. СПб: Изд-во Санкт-Петербургского университета, 1998.
19. *Бондарко Л.В., Вербицкая Л.А., Гордина М.В.* Основы общей фонетики. СПб: Филологический факультет СПбГУ; М.: Издательский центр «Академия», 2004.

Риехакайнен Елена Игоревна

Аспирантка кафедры общего языкознания Факультета филологии и искусств Санкт-Петербургского государственного университета, научный сотрудник Лаборатории экспериментальной фонетики Института филологических исследований СПбГУ. Сфера научных интересов: восприятие речи, изучение спонтанной речи, фонетика, психолингвистика. Автор 20 научных работ.

Основные тенденции развития многоязычной корпусной лингвистики

(Часть 2)



*Потанова Родмонга Кондратьевна,
доктор филологических наук, профессор*

Corpus of Interactional Data (CID) представляет собой УРБД интерактивных аудиовизуальных материалов, собранных и затранскрибированных в Лаборатории языка и речи Университета Экс-ан-Прованса. CID является уникальным источником информации для анализа французской разговорной речи с учётом разных лингвистических уровней (фонетического, просодического, синтаксического, семантического, прагматического и мимико-жестиккулярного).

CID стоит в ряду проектов, связанных с формированием крупных лингвистических БД (и, в частности, УРБД): MATE (Multilevel Annotation, Tools Engineering — Многоуровневое маркирование, технические инструменты); ATLAS (Architecture and Tools for Linguistic Analysis System — Архитектура и инструменты для лингвистических аналитических систем); NITE (Natural Interactivity Tools Engineering — инженерное обеспечение инструментами для естественного общения); Map Task — HCRC (картографирование); DAMSL (Dialog Act Markup in Several Layers — Маркирование диалоговых актов на нескольких уровнях); Verbmobil. Вместе с тем существует очень мало УРБД применительно к французскому языку. Те же, которые существуют (например, COPRAIX, VALIBEL), не всегда доступны. Кроме того, указанные ресурсы создавались для решения сравнительно небольшого круга задач, что накладывает существенные ограничения на возможности их использования для других исследований. Данная ситуация обусловила необходимость в создании такой УРБД, как CID.

Существующие УРБД не являются достаточными для качественного ведения многоуровневого исследования речи. УРБД должна включать записи речи суммарной длительностью в несколько сот часов. Однако лишь небольшая часть этого корпуса сопровождается текстовой репрезентацией. Недавно были предприняты попытки транскрибирования радиотелефонной речи (УРБД ESTER). В других научных областях, таких как лингвистика речевой коммуникации, составляются УРБД всё более сопоставимые по объёму с, например, УРБД CLAPI (Лион). Однако здесь возникает ещё одна трудность: осуществление записи речевого материала в естественной среде (деловой и бытовой дискурс) оказывается делом деликатным. Наконец, процедура собственно анализа фонетико-просодических уровней является чрезвычайно трудоёмкой, поэтому УРБД, содержащие хотя бы тридцать минут звучащего материала, уже могут считаться крупными. Всё это в совокупности объясняет востребованность и необходимость УРБД CID.

УРБД CID создана для удовлетворения информационных потребностей, связанных с разными лингвистическими уровнями: начиная с периферийного (фонетического) и до самого



высокого (дискурсивного, интерактивного), между которыми находятся просодический, синтаксический, семантический, прагматический уровни, а также мимико-жестикуляторного [Bertrand, Blache, Espesser et al. 2006: 31—35].

Поставленная общая задача диктует два, казалось бы, взаимоисключающих требования: обеспечить высокое качество записи, позволяющее проводить анализ на сегментном и супraseгментном (просодическом) уровнях, и в то же время не отбраковывать и не исключать из УРБД диалоги, заслуживающие внимания с учётом уровня коммуникации (например, организации речевых средств, реакции слушателя и т.п.), даже в тех случаях, где качество записи посредственное.

Этапы маркирования и, главным образом, транскрибирования, в силу их трудоёмкости, выполнялись коллективом исследователей. Кроме того, в дальнейшем планируется пополнение УРБД, что обуславливает необходимость в прозрачной и достаточно жёсткой формальной структуре УРБД и столь же прозрачном, формальном и жёстком протоколе записи информации (маркеров) различных уровней.

В настоящее время УРБД CID содержит приблизительно 8 часов диалога на французском языке. Каждый диалог (с участием двух говорящих) длится около 1 часа. Шестнадцать дикторов (10 женщин и 6 мужчин) являются уроженцами разных регионов, однако большинство из них уже в течение многих лет проживают на юго-востоке Франции.

CID представляет собой нечто среднее между многоцелевой УРБД аутентичного речевого материала, подобной CLAPI, и УРБД типа Map Task, которая разрабатывалась под одну конкретную задачу. Собственно, последний тип УРБД лёг в основу множества вариантов УРБД, приспособленных к специфике различных языков (итальянского, шведского, французского и т.п.), а также к решению различных задач.

Дикторы — участники диалогов, включённых в CID, опирались на инструкцию, которая использовалась в качестве инструмента «тематической поддержки», т.е. для того чтобы достаточно быстро завязать разговор и не запинаться в той или иной ситуации. Инструкция не накладывала каких-либо жёстких ограничений на ход диалога, который, таким образом, был реализован максимально естественно.

Запись проводилась в безэховой камере. Участники диалога находились на расстоянии около метра друг от друга, что при обычном разговоре считается естественным. На собеседниках были надеты микрошлемы, позволяющие регистрировать каждый голос на отдельной дорожке. Полученное таким образом оптимальное качество речевого материала позволяет использовать его в других исследованиях. Одно из очевидных преимуществ CID в том, что благодаря раздельной записи голосов места, где реплики накладываются друг на друга, остаются пригодными для всех видов акустического анализа и транскрибирования, что особенно ценно, так как явления, происходящие именно на участках перекрытия реплик диалога, остаются пока малоизученными; в то же время часто звучат утверждения (пока не имеющие подтверждений за отсутствием соответствующего материала), что они играют очень важную роль в структурной организации дискурса.

Кроме аудиозаписи в УРБД CID включена и видеозапись диалогов, что позволяет использовать УРБД для изучения полимодального дискурса.

Некоторое внимание уделялось также психологической подготовке информантов. Участники диалогов должны были быть знакомы с местом записи, чтобы свести к минимуму влияние фактора стресса на естественность речи. Все участники являются сотрудниками (научными сотрудниками или аспирантами) лаборатории, которая занималась формированием УРБД. При делении группы на пары учитывалась степень знакомства информантов и их привычки к общению друг с другом. Наличие привычки гарантирует, что у информантов имеется реальный опыт общения, что способствует большей непринужденности. В этом случае им легче при необходимости отклониться от инструкции и в то же время продолжить следовать всем требованиям эксперимента.

Как уже было сказано выше, все информанты руководствовались инструкцией, в то же время сохраняя естественность речи и позволяя себе свободные импровизации. CID включал в себя разнообразные материалы, в числе которых множество нарративных (главным образом, там, где участники действовали по инструкции), а также доказательных, объяснительных или описательных типов дискурса. Полученные диалоги имеют характер непринужденных бесед: речь может быть достаточно плавной, иногда с перебивками (заполненные паузы, вступления, фальстарты и т.п.). Структурная организация речи подчиняется принципам чередования плавных и неплавных процессов. Наблюдаются так называемые плавные переходы, в которых речь собеседников чередуется достаточно гладко, ритмично, то есть без слишком долгих пауз или «вторжений» в речь собеседника, а также неплавные переходы — многочисленные «перебивы-вторжения» в речь.

Транскрипция в CID, главным образом, орфографическая, опирающаяся на транскрипции GARS. Вместе с тем она а) отражает все типичные явления устной речи, такие как заполненные паузы («ээ-э», «mmm», «хмм» и т.п.), фальстарты, повторы, усеченные слова; б) эксплицитно указывает на имена собственные, географические названия, прямую речь и т.п.; в) содержит некоторые детали фонетического характера (шва, региональная особенность, специфическое произношение и т.п.), необходимые на следующих этапах фонетизации и совмещения с аудиосигналом.

Транскрибирование выполнено с использованием программного пакета PRAAT экспертами-фонетистами. На начальном этапе фонетизации эксперты осуществляли проверку работы программы, чтобы свести к минимуму число возможных ошибок. Перед транскрибированием выполнялось предварительное автоматическое членение речевого сигнала на межпаузальные единицы (ME). ME — блоки речи, ограниченные паузами молчания, как минимум, в 200 мс (длительность может варьироваться в зависимости от ряда факторов). Автоматическая процедура сегментации на ME заключалась в выделении прежде всего глухих / звонких звуков и в установлении пороговой величины для определения паузы определенной длительности. ME часто используется в качестве опорной единицы при формировании УРБД большого объема. По своей формальной и объективной природе эта единица отличается от других просодических единиц, таких, например, как интонационные единицы, членение которых требует ручной работы экспертов, которые, кроме того, могут давать противоречивые интерпретации.

Автоматическая сегментация на ME не только облегчила транскрибирование, но и улучшила эффективность работы на этапах фонетизации и совмещения с аудиосигналом. В тех случаях, когда автоматическая сегментация оказывалась ошибочной, она корректировалась экспертами. На следующем этапе выполнялось автоматическое преобразова-



ние орфографической транскрипции в цепочку символов SAMPA. Для этого использовалась автоматическая программа «Фонетист» (см. табл. 4).

Таблица 4

Фонетико-орфографические соответствия

Орфографическое представление	Фонетическое представление
je_suis	Sui
Allé	Ale
Heu	@
c'est-à-dire	stAdiR
Nourrir	nuRiR@ (южное произношение)

Использованная в ходе работ по формированию CID программа совмещения цепочки символов SAMPA с аудиосигналом была создана в LORIA Д.Фором и И. Лапри. Она опирается на метод HMM (<http://www.loria.fr/equipements/parole/>). В качестве исходной информации эта программа принимает список фонем и аудиосигнал. Результат на выходе — временное местоположение каждой фонемы относительно начала сигнала.

Следует отметить специфические трудности, которые возникали на этапах фонетизации и совмещения. Так, например, например, «je sais» — «я знаю» в беглой разговорной речи иногда произносится как «chai». В подобных случаях решение принималось следующим образом: на этапе транскрибирования и фонетизации отклонения от произносительной нормы передавались транскрипцией. Таким образом, цепочка символов на входе программы оказывалась максимально соответствующей реальному речевому сигналу.

Помимо фонетического и орфографического уровней маркирования в УРБД CID существуют и другие уровни, которые образуют мультимодальную структуру CID.

Уровень сегментного фонетического маркирования имеет (впрочем, как и уровни просодического и мимико-жестикulatoryного маркирования) отличительную черту: единицы этого уровня необходимым образом соотносятся с физической реальностью речевого сигнала. Процесс соотнесения маркеров со звучащим материалом УРБД связан с выбором определённых теоретических и методологических подходов. В противоположность просодическому и мимико-жестикulatoryному уровням, фонетический уровень не ставит проблем, касающихся выбора кода маркирования. В номенклатуре фонетических единиц нет множества возможных теоретических моделей. Зато предметом обсуждения может стать степень точности маркирования (вводить ли какие-то обозначения для переходных процессов артикуляторных фаз и т.п.), а также точности определения позиций пограничных маркеров.

Основная проблема фонетического маркирования состоит во временном определении сегментных единиц речи. Иногда бывает очень сложно найти соответствие между абстрактным и дискретным фонетическим кодом и более или менее непрерывным речевым сигналом.

Одно из явлений, обуславливающих сложность процесса фонетического маркирования, — коартикуляция. Каждая фонема характеризуется совокупностью артикуляторных признаков. Так, для гласного /u/, помимо прочего, характерен такой признак, как огубленность. На артикуляторном уровне этот признак характеризуется выдвиганием губ вперед и их округлением. Данный жест обычно предшествует акустическому эффекту звука, то есть он начинается гораздо раньше, чем произносится сам гласный: в слове /su/ округление губ начинается с момента начала произнесения /s/. Если в приведённом примере представлены два звука в контакте, то было показано, что характерные артикуляторные следы некоторых звуков могут быть идентифицированы на большем расстоянии от самого звука¹. Следствием коартикуляции является то, что временная протяжённость фонетической единицы оказывается достаточно неопределённой и чаще всего превосходит длительность сегмента, который необходимо идентифицировать по речевому сигналу.

Другая сложность связана с отсутствием физических маркеров границ сегментов в потоке речи. В речевом сигнале можно обнаружить много различных «переломных» точек, однако далеко не всегда они соответствуют границам фонетических единиц. Так, одной из таких точек является начало эксплозии взрывного звука или аффрикаты, однако здесь границы сегментов нет. В то же время переход от одного сегмента к следующему может проходить без заметного разрыва, как это имеет место в стечениях гласных звуков. Очевидно, что в этом случае постановка маркера начала/конца сегмента носит произвольный характер и зависит от теоретического и методологического выбора эксперта.

Можно ли в связи с этим утверждать, что любая сегментация речевого сигнала может считаться ошибочной? Вообще говоря — да. Этот вопрос в течение долгого времени вызывал дискуссии в научном сообществе. Необходимо допустить, что разметка границ фонетических единиц — операция, опирающаяся на произвольные критерии и отвечающая необходимости анализа этих единиц. Часто поднимается вопрос о возможности маркирования лишь центра (ядра) квазистационарной части каждой звуковой единицы, что могло бы разрешить проблему границ. Однако такой выбор не позволяет оперировать таким важным параметром, как длительность фонетической единицы, и делает проблематичной синхронизацию многоуровневого маркирования.

При формировании УРБД CID был сделан выбор в пользу постановки маркеров начала и конца каждой единицы сегментного уровня. При этом было сделано допущение о том, что сигнал, заключённый между маркерами, является акустическим соответствием не звука, а лишь части звука. Было показано, что коартикуляция более детально маркируется индексами места образования, в то время как способ образования мог бы служить основанием для более детального временного маркирования. То, что какие-то части сегментов могут при таком подходе оказаться вне участков, ограниченных маркерами, представлялось авторам CID не столь существенным моментом, так как выбранная ими система маркирования позволяла осуществить синхронизацию маркирования на различных уровнях [Bertrand, Blache, Espesser et al. 2006: 37—38].

В спонтанной речи к проблеме границ сегментов добавляется проблема идентификации реально произнесённых фонетических единиц. В ходе работы над УРБД CID эксперты часто транскрибировали сегменты, отсутствующие в сигнале; или, реже, звуки,

¹ Данный признак, связанный с законом антиципации (упреждения и захождения артикуляторных жестов) был описан намного ранее на материале других языков (см., например, Потапова Р.К., Линднер Г. Особенности немецкого произношения. М.: Высшая школа, 1991. 319 с.).



появляющиеся в речевом сигнале, оказывались не записанными экспертами, что не являлось следствием недостаточной квалификации экспертов. Процесс восприятия речи человеком включает реконструкцию отсутствующей или искажённой фонетической информации, и это является его фундаментальным свойством. Реконструкция реализуется вне уровня сознания человека и, таким образом, ускользает от самонаблюдения слушающего. Следовательно, эксперт, выполняющий транскрипцию, может включить в неё и те звуки, которые не были реально произнесены диктором, а только лишь должны были присутствовать в речи. Поэтому после транскрибирования речевого материала УРБД CID проводилась дополнительная коррекция.

В ряде случаев возникали проблемы сегментации спонтанной речи вследствие образования полизвукотипов, особенно на месте цепочек VCV, содержащих звонкий согласный.

Не менее сложной задачей вследствие наличия большого количества параметров, которые необходимо учитывать, является просодическое маркирование. Лишь немногие системы позволяют фиксировать достаточно полную совокупность просодических явлений. Так, TOBI и INSTINT в большей степени ориентированы на интонационные явления. Кроме того, «привязка» к французскому языку такой системы, как TOBI, осложняется её требованиями к априорным знаниям о фонологической структуре исследуемого языка. Напротив, одно из главных преимуществ INSTINT состоит в том, что эта система опирается на акустический анализ, не подразумевающий априори каких-либо знаний о фонологической системе языка. Как утверждают создатели INSTINT, эта система может быть использована в отношении любого языка [Di Cristo et al. 2004].

Что касается именно французского языка, как раз самые последние работы затрагивают вопросы связей между просодическими и дискурсивными явлениями, или же в них делаются попытки описания просодических вариаций, связанных, например, с региональными вариантами, а также предлагаются более полные системы маркирования в виде многоярусной решётки, позволяющие одновременно кодировать феномены общего характера (например, диапазон частоты основного тона) и локального характера (акцентуация, выделение и т.п.) на разных уровнях: временном, метрическом и интонационном. К числу таких систем относится, например, IVTS (адаптированная к системе IViE), которая позволяет фиксировать различные аспекты просодической вариативности.

Система маркирования, учитывающая совокупность просодических феноменов, оказывается необходимой в случаях, когда стоит задача установления связей между различными элементами хотя бы одного просодического уровня, не говоря уже о разных лингвистических уровнях. Выполнение этой задачи применительно к УРБД с объёмом звучащего материала в несколько часов оказывается слишком тяжёлым и дорогостоящим делом в отсутствие комплекса средств, автоматизирующих максимально возможное число этапов кодировки.

При разработке УРБД CID была выбрана система, сочетающая ручной и автоматический методы обработки, которая опирается в первом случае на слухо-

вую идентификацию экспертами частных просодических феноменов (подход, сходный с TOBI) и на подход MOMEL-INSTINT во втором случае [Hirst et al. 2000].

По этой причине при маркировании CID были использованы несколько видов маркеров. Во-первых, применялась нотация INSTINT, которая позволяет автоматически кодировать тональные целевые сегменты. Система INSTINT использует алгоритм MOMEL, позволяющий моделировать кривую F_0 , и даёт на выходе последовательность лингвистически релевантных точечных целей. Система INSTINT имеет алфавит из восьми символов: Top, Middle и Bottom определяются в целом, по отношению к регистру каждого говорящего; Higher, Same и Lower определяются по отношению к предыдущим позициям, как и Downstepped и Upstepped, относящиеся к более слабым изменениям.

Другой набор знаков применялся при ручном маркировании. Один из них связан с просодической фразировкой высказываний, то есть с определением областей просодических единиц. Выделялись интонационные единицы (*les unités intonatives*) и единицы акцентуации (*les unités accentuelles*). Специальный маркер был введён для неоднозначных случаев или случаев, которые невозможно отнести к первой или второй категориям. Подобные случаи могут быть связаны с присутствием дискурсивных маркеров (таких как «что?» (*quoi*), «видишь ли» (*tu vois*) и т.п.).

Ряд маркеров относился к «интонационным контурам». Были приняты следующие символы: ровный / flat (fl), малый подъём / minor rising (mr); другие малые контуры / other minors (m0); нисходящие / falling (F); восходяще-нисходящие / rising-falling (RF1); восходяще-нисходящие с предпоследнего слога / rising-falling from penultimate (RF2); большой подъём, предполагающий продолжение / major continuation rising (MCR); терминальный подъём / terminal rising (TR); вопросительный подъём / question rising (QR); подъём при перечислении / enumerative rising (ER); падение при перечислении / enumerative falling (EF).

Принцип морфосинтаксического маркирования заключался в присвоении словам высказывания соответствующих категорий. Существует несколько систем, или классификаторов, позволяющих с большим или меньшим успехом автоматически выполнять эту задачу.

В частности, для французского языка известны такие системы, как WinBrill, Cordial и LPL. Последняя была применена в ходе работы над УРБД CID. Она использует стохастические данные, полученные при отработке обучающей выборки, для того чтобы определить, какие морфосинтаксические маркеры наиболее вероятны для того или иного высказывания. В настоящее время авторский коллектив CID работает над адаптацией классификатора LPL к задаче пополнения УРБД [Bertrand, Blache, Espesser, 2006].

Синтаксическое маркирование остаётся сложной задачей, с трудом поддающейся автоматизации. Несмотря на это, существует ряд автоматических анализаторов, которые могут использоваться, как минимум, в качестве основы для осуществления маркирования. В связи с этим, в зависимости от желаемого уровня описания, различают подходы двух типов. Самое простое маркирование (проставление скобок) может опираться только на методы поверхностного анализа (*shallow parsing*). Тем не менее, возможно также более тонкое маркирование, выполняемое с помощью более сложных анализаторов. Последние, в отличие от поверхностных анализаторов, позволяют идентифицировать не только единицы и их структуры, но также и синтаксические отношения, связывающие их одновременно с грамматическими функциями этих единиц. Для всех случаев (эта ремарка применима для всех представленных здесь уровней анализа) техника маркирования и используемые формальные приёмы независимы от теоретического



подхода, выбранного для синтаксического описания (например, HPSG, GP или грамматики соподчинённости).

Поверхностные анализаторы предоставляют информацию о границах составляющих текста. Этот тип анализа используется также в более широких областях: например, в области передачи информации, диалоговых систем, систем синтеза речи. Данный тип программы опирается на совокупность правил, определяющих левые и правые границы составляющих в зависимости от анализируемой составляющей и различных свойств читаемого слова.

С недавнего времени существуют несколько средств запроса к синтаксически маркированным УРБД как результат усилий, направленных на формирование синтаксически маркированных французских баз данных. Средства для формирования запросов в синтаксически маркированных УРБД позволяют учитывать соотношения по принципу вычленения основной информации.

Данное маркирование разработано в настоящий момент для уровня лексики и отношений между лексическими единицами. Речь идёт об упорядочивании элементов, релевантных для построения смысла дискурса, и одновременно о маркировке лексических единиц и связывающих их отношений. Составление подобного массива — первый шаг в формализации семантических и дискурсивных связей.

Семантическое маркирование, целью которого является упорядочение лексических и межлексических единиц, должно включать, как минимум, три информационных уровня:

— на первом уровне маркирования отмечаются семантические функции. Под этим подразумевается точная маркировка вклада лексической единицы (например, состояние, процесс и переход) в упорядочение событийной структуры фразы, которая позволяет отобразить особый уровень информации в лексической семантике. Если точнее, то она предназначена для классификации объектов мира. Структура имеет четыре функции, уточняющие семантические признаки единицы информации: конститутивные, формальные, целевые и агентивные. В соответствии с этим уточняется отношение объекта к его составляющим, характерные признаки объекта, функции объекта и активных участников, связанных с объектом.

На втором уровне отмечается информация онтологического типа. Эта совокупность семантических черт отбирается на основе предварительных лингвистических исследований, доказавших свою эффективность во многих языках.

Наконец, на третьем уровне отмечаются отношения между лексическими единицами. Речь идёт, таким образом, одновременно о фиксации отношений иерархического порядка (например, анафорических), но также (и в основном) об указании на то, как был получен смысл каждой единицы высказывания, что может выполняться с учётом взаимодействий той или иной единицы с другими полисемическими единицами высказывания. Поскольку семантическая единица формируется только в контексте, необходимо отметить вклад каждого уровня в формирование смысла.

Для того чтобы вести речь о прагматическом уровне, необходимо уточнить содержание этого термина. В литературе он ассоциируется с различными теоретическими направлениями, методологиями, с множеством самых различных объектов исследования. В данном случае этот термин охватывает три перспективы, определяющие уровни разного характера: например, языковые действия, феномены разговорного порядка, связанные с конструированием речевых оборотов, и другие уровни, также релевантные в нарративном аспекте.

Ниже приведён пример маркирования. Первый элемент, имеющий временной образец, служит индексом для других элементов, которым самим присваивается индекс по их позиции в структурном элементе, в который они входят.

```
<el индекс=«32» начало= «5.8588» конец= «6.0908»>
<имя-атрибут= «SpellSp1»>графемы</атрибут></el...
  <el индекс=«26» начало= «32» конец= «32»>
<имя-атрибут= «Имя Нарисательное»>Нарисательное</ атрибут >
  <имя-атрибут=Согласие>4</атрибут></el>...
  <el индекс=«15» начало= «31» конец= «32»>
<имя-атрибут= «Именное предложение «>Стандарт</атрибут>
  <имя-атрибут=Согласие>4</атрибут></el>...
```

Промаркированы также и другие типичные элементы устной речи. Первая линия маркирования относится к дискурсивным маркерам типа «что», «вот», «видишь ли», «знаешь ли», «наконец». Две другие линии маркирования касаются слуховых феноменов. Среди слуховых сигналов различаются простые и сложные сигналы: некоторые повторы, повторные формулировки, метапросы, завершения и т.п. Одна линия маркирования посвящена формальным категориям («ммм», «ну да», «а, нет», «а, ну да», «согласен» и т.п.), другая линия посвящена функциональным категориям. Подразумевается, что деление на сложные категории уже включает функциональный аспект (минимальное выслушивание, взятие на заметку, оценка, суждение).

Последняя линия маркирования относится к типу единиц, используемых для учёта речевых оборотов: единицы построения оборотов речи / turn-constructional units (TCUs). TCUs — это единицы, считающиеся потенциально «завершёнными» с синтаксической, просодической и прагматической точки зрения. Конец TCU представляет собой место потенциального завершения, отсылающее к так называемому переходному месту / transition-relevance place (TRP).

TCUs являются, таким образом, наименьшими лингвистически незавершёнными и завершёнными единицами, релевантными на уровне коммуникации. Конечная TCU представляет собой оборот, состоящий из одной единицы, в то время как неконечная TCU — один из семантически или прагматически незавершённых компонентов сложного оборота, определённого, например, в терминах дискурсивной деятельности (каузальные конструкции, повествовательная последовательность и т.п.). CID маркирован с учётом конечных и неконечных TCU.

Если, в отличие от фонетического маркирования, совмещение дискурсивной деятельности с речевым сигналом и не является абсолютной необходимостью, то прагматическое маркирование подобных языковых феноменов ставит, тем не менее, ряд проблем. Первая проблема заключается в понимании самого термина «маркирование». Вторая проблема — в принятии во внимание фактора временной развёртки речевого высказывания, а следовательно, и границ наблюдаемого явления.



Размытость дискурсивных и нарративных границ и их контекстуальный фон (принятие в расчёт ситуации общения, скрытые обстоятельства, реакции говорящих и т.д.) делают вопрос маркирования проблематичным. Становится необходимым работать по возможности с наиболее полным источником информации, включая видеоряд. Полимодальный анализ на данном уровне имеет, следовательно, первостепенное значение, по крайней мере, по двум причинам: первая касается интерпретации исследуемых явлений, которая может быть лишь улучшена благодаря учёту всех прочих уровней, участвующих в формировании смысла, вторая связана с совершенствованием и обогащением процессов маркирования.

Невербальное маркирование УРБД находится в процессе разработки. В рамках исследования было использовано программное обеспечение ANVIL, учитывающее ручную жестикуляцию и выражение лица, движения головы, направление взгляда. Этот код был дополнен и адаптирован в соответствии с потребностями исследования (в частности, были введены обозначения для движений корпуса, а также уточнены эталоны для дейктических жестов) [Bertrand, Blache, Espesser 2006: 50—51].

Жесты, включённые в номенклатуру маркеров, перечислены в таблице 5.

Таблица 5

Перечень жестов, маркированных в системе CID

Голова / Лицо	Руки	Корпус
Выражение лица	Симметрия / асимметрия жеста	Движения корпуса
Движение бровей	Траектория руки	
Открытие глаз	Конфигурация руки	
Направление взгляда	Семиотический тип жеста	
Открытие рта	Фазы жеста	
Конфигурация губ	Вершина	
Движения головы	Точка контакта	
Эмоции	Высота реализации жеста	
	Позиция жеста в пространстве жестикуляции говорящего	

Жест включает разные фазы, такие как фаза подготовки (рука, например, покидает положение покоя и вступает в реализацию жеста), фаза собственно реализации жеста, затем фаза исхода (когда рука для того, чтобы возобновить тот же тип жеста, возвращается в позицию покоя). Перед конечной фазой жест может быть задержан. К этим фазам добавлен параметр, отмечающий экстремум жеста (точку, когда жест достигает своего максимального развёртывания по отношению к положению покоя). Наконец, точка контакта используется для адаптивных жестов, фиксирующих контакт между рукой и какой-либо частью корпуса либо говорящего, либо партнёра по коммуникации.

В настоящее время на основе этих стандартов создан файл (см. листинг 1). В приведённом фрагменте указано, что линия маркирования Eyebrows (движения бровями) входит в группу маркирования движений лица (group name=Face) и что для движений бровями можно встретить такие значения, как frowning (нахмуривание бровей) или raising (поднятие бровей). Введено также значение other (другое) для редких случаев, когда говорящий может, например, поднять одну бровь.

Листинг 1

<pre><?xml version="1.0" кодировка="ISO-8859-1"?> <annotation-spec> — <head> — <valuetype-def> — <valueset name="EyebrowsType"> <value-el color="#9df4a">Frowning</value-el> <value-el color="#f1f07a">Raising</value-el> <value-el color="#f5ce16">Other</value-el> </valueset> </valuetype-def> </head> — <body> — <group name="Face"> — <track-spec name="Eyebrows" type="primary"> <attribute name="Eyebrows" valuetype="Eyebrows type" display="true"/> </track-spec> </group> </body> </annotation-spec></pre>	<p>Данная первая строка описывает файл как файл xml Наименование маркирования В первой части файла даны все возможные в линии значения маркирования, здесь — <i>frowning, raising, other</i></p> <p>Во второй части файла описана линия <i>eyebrows</i> и указано к тому же, что линия эта первична (она не зависит от других линий), но входит в состав группы <i>face</i>, группы маркирования движений лица. Синтаксис XML требует, чтобы все скобки были закрыты.</p>
---	---

Создание файла позволило, в частности, продумать иначе структуру маркирования жестов: так, вместо того, чтобы создавать множество специфических маркеров, таких как:

Таблица 6

Примеры маркеров

Gaze >	Sideways >	left / right
Head >	Side turn >	left / right
	Single tilt >	left / right
Trunk >	Sideways >	left / right
Hands >	Single hand >	left / right

— достаточно ввести один раз значение left / right, применимое к движениям любых частей тела. Если движение не предполагает параметра left / right, как например, eyebrow raising / frowning, то к ярлыку left / right по умолчанию применяется значение none. Таким же образом функционирует маркер concrete / abstract, который применим только к дейктическим жестам и направлению взгляда. Параметр же contact point относится только к приспособительным жестам, но, однако, присутствует и отмечен как none для всех других жестов.

Как только маркеры вводятся в ANVIL, программное обеспечение порождает файл XML, указывающий для каждой совокупности маркеров ранг, время начала и время конца.



Значения по умолчанию не учитываются в заключительном файле маркеров и, следовательно, его не загромождают ненужной для того или иного типа жеста информацией.

Помимо маркирования языка жестов, ANVIL позволяет группировать описанные выше совокупности маркеров других уровней. Данное программное средство позволяет не только импортировать совокупности маркеров, например, из PRAAT, но также их менять при условии предварительной спецификации ярлыков, используемых в спецификационном файле. Это привело к необходимости создания иерархии маркеров в рамках каждого уровня: например, последовательности маркеров, представленные в линейном виде в формате PRAAT, были построены в соответствии с иерархией, отвечающей требованиям структуры формата XML, входного и выходного формата ANVIL. Что касается других уровней, маркирование которых не выполнялось в PRAAT, то они выполнены в формате XML, при этом маркеры также были специфицированы в файле *sres*.

Следует отметить, что формат ANVIL предусматривает возможность создания столько файлов *sres*, сколько требуется, используя тот или иной уровень маркирования, а также тот или иной набор тэгов. Возможна также последующая непосредственная модификация этих файлов, поскольку они создаются в формате XML.

Интерес к применению комплексного подхода в маркировании УРБД (в частности, многомодальных УРБД) основан на перспективе самых широких областей применения этих данных. В частности, предусматривается возможность запросов высокого уровня с привлечением различных областей маркирования, позволяющих более систематично исследовать типы взаимодействия, которые могут иметь место между ними. Необходима временная синхронизация, если маркированные объекты четко идентифицированы и отсегментированы. Вопрос синхронизации — это главная проблема мультимодальных УРБД: как установить связь между объектами, принадлежащими к разным областям, без единой справочной базы? В ходе работы над УРБД CID было выбрано решение, опирающееся на позиционирование каждого объекта с помощью специальной системы. Данный принцип заключается в указании, если это возможно, для каждого объекта нескольких ориентиров. Разумеется, большинство объектов, напрямую связанных с речевым сигналом, имеют соотношение с временной позицией. Равным образом, как указано выше, некоторые объекты из других областей могут быть также «привязаны» к временному сигналу, что касается, например, слов и морфосинтаксических конструкций. В этом случае может быть предложен также другой тип привязки: «позиция» объекта в цепи (например, ранг слова в последовательности или фразе). Такой тип более традиционен для обработки письменного материала, но в данном случае он применим и к устно-речевому дискурсу. Наконец, для включения в комплекс семантических данных предложено использование индексации элементов речи. В целом, ориентировка каждого элемента может проводиться через комплексную «привязку», поясняющую все три составляющие (временную реализацию сигнала, положение в речевой цепи, индекс контекста) или их часть.

После того как этот процесс закончен, становятся возможными запросы в УРБД с использованием различных уровней маркирования. В этом случае речь

идёт о запросах одновременно по нескольким областям при синхронизации поисков через комплексную систему «привязок». Следующие примеры запроса иллюстрируют функционирование системы:

Q0 Найти паузы внутри синтаксической единицы

Требуемые области — синтаксис и просодия. Поиск заключается в идентификации объектов особого просодического типа, границы которых включены в границы объекта, принадлежащего к другой области — синтаксису. В этом случае, временные и позиционные «привязки» позволят идентифицировать объекты.

Q1 Найти дейктические жесты, ассоциируемые с местоимением

Данный запрос относится одновременно к области жестов и морфосинтаксической области. Он заключается в идентификации в одной области (жестов) всех объектов особого типа (дейктических) и выявлении перекрытия этой области объектами типа «местоимение» в определённых пределах (интервалы которых включены в пределы интервалов объектов, относящихся к жестам).

Q2 Найти синтаксические субъекты

В данном случае отобранные объекты должны иметь синтаксическую характеристику (быть фрагментами), семантическое свойство (быть субъектами) и принадлежать к особому просодическому контуру (например, с восходящей мелодикой). Здесь полезны «привязки» временные и позиционные.

Q3 Найти каналы обработки связи

Цель данного запроса — идентификация особых речевых маркеров, реализуемых говорящим и слушающим.

Возможность осуществления подобных запросов позволяет, таким образом, весьма полно использовать ресурсы мультимодальной УРБД. Такой тип исследования в УРБД большого объёма на сегодня не имеет аналогов, которые содержали бы достаточно полное многоуровневое маркирование. Подход, применённый в УРБД CID, представляет собой эффективное решение, позволяющее синхронно отображать маркеры различных уровней. Таким образом, речь идёт о важном вкладе в анализ мультимодальных систем.

Рост числа лингвистических работ, посвящённых просодии, речи и грамматике, с одной стороны, а также развитие исследовательских подходов к УРБД, с другой стороны, побуждает к размышлениям о символическом и дискретном отображении непрерывных просодических элементов. Вопрос кодировки, символического и дискретного отображения супrasegmentных единиц не является нейтральным, особенно если им задаваться с целью установления параллели с отображением сегментного уровня, предложенным IPA (International Phonetic Association и International Phonetic Alphabet — Международная фонетическая ассоциация и международный фонетический алфавит). Данная система кодировки опирается на несколько гипотез, касающихся континуума речи и его анализа, а именно: а) одни аспекты речи лингвистически релевантны, в то время как другие — нет. Это затрагивает вопрос о возможности выполнения широких (или фонематических) транскрипций и узких (или фонетических) транскрипций; б) континуум речи может быть частично отображён как последовательность сегментов.

Важно отметить, что единственное общеизвестное допущение, которое делает IPA, — это то, что континуум речи может быть отображён в форме дискретных сегментов. Однако использование чёткого символа для отображения звука или сегмента не обязательно превращает последний в фонему.



Один из источников ошибок связан с трудностями задачи транскрибирования. Действительно, транскриптор может по-своему кодировать услышанные звуки, делать выбор, объясняемый либо знанием фонологической системы транскрибируемого языка, либо незнанием этого языка. В некоторых случаях он может выбрать транскрипцию сегментов не в том виде, в каком они были действительно произнесены, а как образцы фонем. Следовательно, для разработки или оценки любой системы транскрипции просодических единиц необходимо ввести систему ограничений и провести сопоставление четырёх систем транскрибирования просодических единиц: IPA и специфические символы кодировки супraseгментных единиц, INSTANT, ToBI и IVTS — систему транскрипции, разработанную на основе IViE. При этом ставится двойная цель:

- представить каждую систему транскрипции и
- оценить способы, с помощью которых эти различные системы могут транскрибировать некоторые типы особых просодических явлений, в частности, акцентуацию, мелодику и тональность [Delais — Roussarie, Post, Portes 2006: 63—65].

Как и ToBI, система IVTS использует несколько категорий маркирования для кодировки различных просодических и лингвистических данных. В IVTS транскрипция организована на шести уровнях, из которых четыре используются для записи просодических феноменов. Транскрипция, таким образом, принимает следующую форму:

Таблица 7

Уровни транскрибирования в системе IVTS

{ Категория «Комментарии» (или <i>Comments tier</i>)
{ Категория «Фонология» (или <i>Phonological tier</i>)
{ Категория «Общее фонетическое восприятие»
Декодирование просодической информации
(или <i>Global auditory phonetic tier</i>)
{ Категория «Локальное фонетическое восприятие»
(или <i>Local auditory phonetic tier</i>)
{ Категория «Ритм» (или <i>Rhythmic tier</i>)
{ Категория «Слово» (или <i>Orthographic tier</i>)

Каждый из уровней служит для кодировки отдельных данных. Категория «Слово» используется для совмещения произнесённых слов с участками сигналов, которые им соответствуют. В других категориях обозначения совмещаются с определёнными точками сигнала, например:

- а) участком слога, воспринимаемого как выделенный;
- б) границами интонационных областей и т.п.

В категории «Комментарии» точки, взятые для совмещения с сигналом, соответствуют зонам, к которым относятся комментарии. Различные виды принятых совмещений зависят от их релевантности для обрабатываемого языка и от характера работы привязки мелодических контуров.

В категории «Ритм» обозначение R указывает на то, что отмеченный слог выделен сильнее, чем соседние слоги. На акустическом уровне это

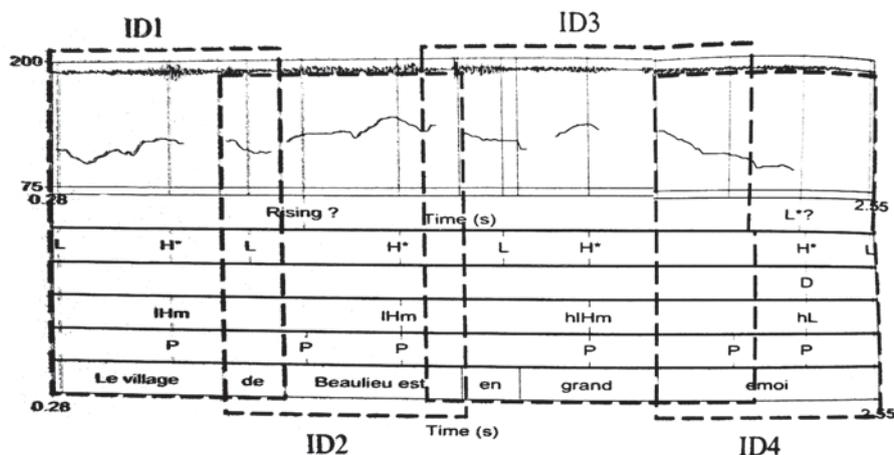


Рис. 4. Сегментация с учётом реализации акцентуации (ID) фразы «le village de Beaulieu est en grand émoi» (деревня Больё в большом волнении)

может характеризоваться увеличением длительности, мелодическим контуром и т.п. Отметим, что P указывает на воспринимаемое качество, но не обязательно на абстрактное структурное свойство слова или группы слов с лексическим ударением. Кроме того, обозначения P присваиваются слогам, на которых реализуются особые мелодические движения, — их предстоит совместить с обозначениями категорий «Локальное фонетическое восприятие» и «Фонология». При этом последняя операция не является обязательной.

Категория «Локальное фонетическое восприятие» используется для записи формы мелодических движений, выполненных на выделенных слогах, а также на примыкающих к ним слогах (см. рис. 4).

Здесь акцент делается на мелодической конфигурации и характере совмещения отдельных участков. Мелодические изменения более общего плана, такие как регистр или нисходящая мелодика, на данном уровне не кодируются. Транскрипция мелодических изменений проводится на перцептивно-слуховой базе, а не на основе акустического анализа частоты основного тона. Она выполняется при внимательном прослушивании части сигнала, соответствующего области реализации акцентуации (ID). Протяжённость этой области варьирует в зависимости от языков. Во французском языке любая ID включает а) выделенный слог, обозначенный P; б) все предшествующие ему слоги до предыдущего выделенного слога или до границы основной интонационной области; в) непосредственно следующий за ним слог. Согласно этому определению высказывание «le village de Beaulieu est en grand émoi» (деревня Больё в большом волнении) делится на четыре ID, по одной на каждый выделенный слог, отмеченный мелодическим движением.

Вначале система транскрипции IPA была разработана для кодирования сегментной информации. При этом был предложен набор символов для кодировки некоторых просодических феноменов как метрического, так и тонального характера. Для кодировки метрических феноменов IPA предлагает два различных символа, «'» и «,»; первый для отображения слогов, получающих первичное ударение, второй для слогов, получающих вторичное ударение. Другая серия используется для отображения просодических при-



знаков. Приняты два уровня структуризации: основание (или малая группа), представленная символом « | », и интонационная (большая) группа, представленная символом « || ». Для феноменов тонального характера в IPA имеются две серии символов: одна для статических тонов, другая для модулированных тонов. Эти символы созданы для транскрибирования лексических тонов в таких языках, как китайский; однако они не позволяют выполнять кодировку интонационных фразовых явлений. В этом случае могут быть использованы некоторые символы широкого значения: символ нисходящей шкалы \$ и восходящей шкалы #, символы нисходящей мелодики (и восходящей мелодики &.

Для маркирования ударных слогов IPA предлагает фиксировать различие между первичным и вторичным ударениями, то есть привлекать фонологический уровень. Такой подход предполагает, что ритмическое функционирование языка известно, а значит, можно определить, является ли физическая выделенность первичным или вторичным ударением. Следовательно, на сегментном уровне возможности, связанной с различием между широкой и узкой транскрипцией, здесь, по-видимому, не существует.

Вторая трудность соотносится с применением символов сегментации на просодические группы. IPA предлагает установить различие между двумя уровнями просодической структуризации: малой группой и интонационной группой. При этом ничего не говорится о критериях определения этих групп [Delais Roussarie et al. 2006].

Изучение различных примеров в Handbook of the International Phonetic Association (1999) показывает, что дело обстоит отнюдь не так гладко. В каталанском языке выбор между двумя уровнями кажется подчинённым использованию или неиспользованию терминального контура: если последовательность завершается продолжением, то граница между малыми группами используется, в противном случае используется граница между большими группами. В транскрипции французского языка границы между большими группами используются как после контура продолжения, так и после финального контура. Следовательно, выбор между тем или другим символом оказывается часто делом тонким и зависит от транскриптора. Отметим, впрочем, что эта проблема сегментации существует постоянно в большинстве систем транскрипции, за исключением, возможно, ToBI и break indices, образующих более прочный и надёжный инвентарь.

Институт фонетики Экс-ан-Прованса (Франция) предлагает теоретическую модель и инструменты маркирования интонационных систем. Оригинальность подхода заключается в том, что предложено средство автоматического маркирования на двух уровнях: уровне «фонетической» репрезентации, порождённой алгоритмом MOMEL (Modeling MELody), и уровне «поверхностно-фонологической» репрезентации, автоматически выполняемой с помощью алфавита INSTINT (International Transcription System for INTonation).

Алгоритм MOMEL и «фонетическая» репрезентация частоты основного тона преобразуют прерывистую кривую — следствие огрублённого распозна-

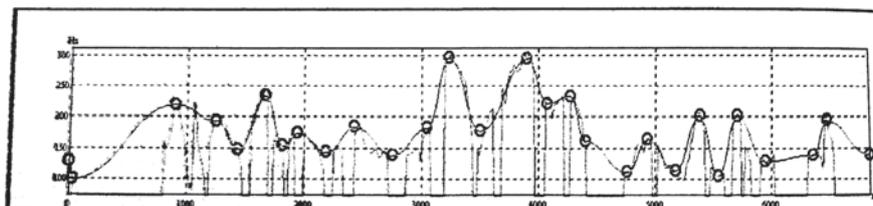


Рис. 5. «Целевые точки» и сплошная кривая, порождённая алгоритмом MOMEL на основе выделения частоты основного тона

вания частоты основного тона — в сплошную кривую, которая является интонационно релевантной. Роль алгоритма заключается в отделении «макропросодической» составляющей от «микросодической» составляющей, которая не принимается во внимание как лингвистически нерелевантная. На выходе MOMEL производит совокупность точек, охарактеризованных с помощью пары «временная локализация / F_0 ». Эти точки затем объединяются. Зоны отклонения представляются вершинами и впадинами.

Алфавит INSTINT включает набор из восьми абстрактных тональных символов. Три из этих символов кодируют «абсолютные» тоны, разграничивающие общую протяжённость регистра говорящего на отрезке «интонационной единицы» (соответствующей максимальной единице просодической фразы для данной модели): речь идёт о символах T для Top, M для Mid и B для Bottom. Пять остальных символов кодируют «относительные» тоны, значение которых зависит от значения предыдущего тона. Относительные тоны делятся на две подкатегории: неитеративные тоны (H — для высоких, S — для монотонных и L — для низких) и итеративные тоны (U — для повышения и D — для понижения). Помимо орфографических символов второй набор диакритики используется преимущественно в рамках транскрипции текстов.

Возможно получение автоматической кодировки речевых данных INSTINT с помощью алгоритма, учитывающего два дополнительных параметра: ключ и регистр. И тот, и другой зависят одновременно от говорящего и высказывания. Абсолютные тона T и B определены как границы тонального регистра говорящего, симметрично распределённого вокруг ключа, характеризующего значение тона M.

* * *

В заключение следует подчеркнуть, что языковые ресурсы являются важнейшим компонентом процесса создания и эксплуатации различного рода информационных систем, реализующих лингвистические функции, направленные на обработку естественного языка в его различных проявлениях (применительно к печатным и рукописным текстам, а также к звучащей речи).

В области корпусной лингвистики современные компьютерные технологии ускоряют и упрощают процедуры лингвистической обработки больших массивов текстов в их письменном и устном вариантах. По сути, лингвистический корпус — это своего рода информационно-справочная система, основанная на текстовых ресурсах на некотором языке в электронной форме при наличии особой дополнительной информации о свойствах лингвистического материала.



Таким образом, корпусная лингвистика — это бурно развивающаяся отрасль «лингвистической индустрии», предназначенная как для проведения научных исследований, так и для решения целого ряда прикладных задач.

Литература

1. Автоматизированное рабочее место эксперта-фоноскописта. Электронная энциклопедия, версия V1.0: <http://www.estra.ru>
2. Андрющенко В.М. Концепция и архитектура Машинного фонда русского языка. М., 1989.
3. Белолипецкий С.И., Буря А.Г. Специализированные СУБД для поддержки речевых баз данных // Сетевой электронный научный журнал «Системотехника». №2. 2004. М.: МГИЭМ, 2004.
4. Богданов Д.С., Брухтий А.В., Кривнова О.Ф., Подрабинович А.Я., Строкин Г.С. Технология формирования речевых баз данных // В сб. «Организационное управление и искусственный интеллект». М.: Эдиториал УРСС, 2003.
5. Богданов Д.С., Кривнова О.Ф., Подрабинович А.Я., Фарсобина В.В. База речевых фрагментов русского языка ISABASE // В сб.: «Интеллектуальные технологии ввода и обработки информации». М., Эдиториал УРСС, 1998.
6. Богуславский И.М., Григорьев Н.В. и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации // Труды Международного семинара ДИАЛОГ-2000. М., 2000. Т. 2. С. 41—47.
7. Корпусная лингвистика в России. / Сост. Е.В. Рахилина и С.А. Шаров // Спец. выпуск журнала НТИ. М., 2003. Сер.2. № 6, 10.
8. Кривнова О.Ф., Захаров Л.М., Строкин Г.С. Речевые корпуса (опыт разработки и использование). М.: МГУ. [http://www.dialog-21.ru/archive_article.asp].
9. Кривнова О.Ф., Захаров Л.М., Строкин Г.С. Речевые корпуса (опыт разработки и использование) // Сборник трудов Международного семинара Диалог'2001 по компьютерной лингвистике и её приложениям (в двух томах). Т. 2. Прикладные проблемы. М., 2001.
10. Кривнова О.Ф. Области применения речевых корпусов и опыт их разработки // Сборник трудов XVIII сессии РАО. М.: ГЕОС, 2006.
11. Леонтьева Н.Н. Автоматическое понимание текстов: Системы, модели, ресурсы. М.: Академия/Academia, 2006.
12. Потапова Р.К., Линднер Г. Особенности немецкого произношения. М.: Высшая школа, 1991. 319 с.
13. Потапова Р.К. Лингвистическое обеспечение Электронной Энциклопедии, предназначенной для экспертов-фоноскопистов (русский язык). М.: ЭСТРА, CDROM, 1998–1999.
14. Потапова Р.К. Новые информационные технологии и лингвистика. 4-е изд., суц. доп. — М.: Эдиториал УРСС, 2005. — 368 с.
15. Потапова Р.К. Речь: коммуникация, информация, кибернетика. М.: Радио и Связь, 1997. 528 с.
16. Потапова Р.К.. Тайна современного кентавра. Радио и связь, М., 1992.
17. Рыков В.В. Корпус текстов — новый тип словесного единства // Труды Международного семинара ДИАЛОГ-2003. Протвино, 2003.
18. Сичинава Д.В. К задаче создания корпусов русского языка в Интернете // НТИ. М., 2002. Сер.2. № 12.

19. *Скрелин П.А., Щербаков П.П.* Требования к современной фонетической базе данных для фундаментальных и прикладных исследований // Технологии информационного общества — Интернет и современное общество: труды VI Всероссийской объединенной конференции. Санкт-Петербург, 3—6 ноября 2003 г. СПб.: Изд-во Филологического ф-та СПбГУ, 2003. С. 62—63.
20. *Шаров С.А.* Параметры описания текстов корпуса, а также Корпусная лингвистика в России // НТИ. М., 2003. Сер.2. № 5—6.
21. *Arlazarov V.L., Bogdanov D.S. Krivnova O.F., Podrabinovitch A.Ya.* Creation of Russian Speech Databases: Design, Processing, Development Tools. // International Conference SPECOM'2004. Proceedings. S-Pb. Russia, 2004. Pp: 650—656.
22. *Barlow M.* Corpora for Theory and Practice. //IJCL. Amsterdam, 1996. № 1.
23. *Bel B., Blache P.* Le Centre de Ressource pour la Description de l'Oral. // Corpus. Les enjeux de l'annotation. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence, 2006. Pp.13—18.
24. *Bertrand R., Blache P., Espesser R., Ferre G., Meunier C., Priego-Valverde B., Rauzy S.* Le CID — Corpus of Interactional Data: Protocoles, Conventions, Annotations. // Corpus. Les enjeux de l'annotation. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence, 2006. Pp.31—60.
25. *Bohmova A.* Automatic Procedures in Tectogrammatical Tagging. //The Prague Bulletin of Mathematical Linguistics. Prague, 2001. №76. P.23—34.
26. *Collier A., Pace y M., Renouf A.* Refining the Automatic Identification of Conceptual Relations in Large-scale Corpora. // Proceedings of the Sixth Workshop on Very Large Corpora. Montreal, 1998.
27. *Delais — Roussarie E., Post B., Portes C.* Annotation prosodique et typologia. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence. Vol.25, 2006. P. 61—95.
28. *Greenstette G., Segond F.* Multilingual Natural Language Processing // IJCL. 1997. V.2. № 1.
29. *Hajicovd E., Pajas P., Vesela K.* Corpus Annotation on the Tectogrammatical Layer: Summarizing of the First Stages of Evaluations // The Prague Bulletin of Mathematical Linguistics. Prague, 2002. №77. P. 5—18.
30. International Journal of Corpus Linguistics (IJCL). / Ed. W.Teubert. — Amsterdam, 1996—2001.
31. *Kibkalo A.A., Lotkov M.M.* Choice of Phonetic Alphabet for Russian LVCSR System // Proceedings of the International Workshop «Speech and Computer» SPECOM' 2003. (Moscow, 27—29 October, 2003) Moscow: MSLU, 2003. P. 102—105.
32. *Kucova L., Hajic ova E.* Prague Dependency Treebank: Enrichment of the Underlying Syntactic Annotation by Coreferential Mark-Up // The Prague Bulletin of Mathematical Linguistics. Prague, 2004. №81. P. 23—34.
33. *Lee Y.-J., Choi D.-L., Um Y., Lee K.-H., Kim Y.-I., Kim B.-W.* Speech Resources at SITEC in Korea // Proceedings of the 10th International Conference SPEECH and COMPUTER (SPECOM' 2005) (Patras, Greece, 17—19 October, 2005) Patras, Moscow: MSLU, 2005. P. 579—582.
34. *Loseva E., Potapova R.* Speech variability of vibrants: phonetic database for English and German // Proceedings of the 10th International Conference Speech and Computer SPECOM' 2005, Patras, Moscow: MSLU, 2005.
35. *Marcus M.P., Santorini B., Marcinkiewicz M.A.* Building a Large Annotated Corpus of English: The Penn Treebank // Computational Linguistics. 1993. Vol.19. №2. P. 313— 30.
36. *Potapova R.K., Potapov V.V.* Database of forensic phonetics knowledges (as applied to electronic encyclopaedia for Russian experts) // Proceedings of the International Conference of IAFP, York, UK, 1999. P. 6—7.
37. *Shaikevich A.* The Computer Fund of Russian Language // IJCL.-Amsterdam, 1997. V.2. №1. P. 163—167.
38. *Teubert W.* Corpus Linguistics and Lexicography // IJCL. Philadelphia, 2001.
39. http://www.mdi.ru/asnews/body/03.12.2001_39303.html
40. <http://cfrl.ru>
41. <http://conf.infosoc.ru/03-r2f14.html>



42. <http://www.auditech.ru>
43. <http://www.auditech.ru>
44. http://www.mdi.ru/aspnews/body/03.12.2001_39303.html

Родмонга Кондратьевна Потапова

Академик Международной академии информатизации, доктор филол. наук, профессор. Заслуженный работник Высшей школы РФ.

Зав. отделением прикладной лингвистики, зав. кафедрой прикладной и экспериментальной лингвистики, директор Центра фундаментального и прикладного речеведения Московского государственного лингвистического университета. Специалист в области романо-германского языкознания, общей и прикладной фонетики, теоретической, прикладной, экспериментальной и математической лингвистики. Автор свыше 450 научных и научно-методических публикаций.