УДК 621.372 : 519.72

АВТОМАТИЧЕСКИЙ АНАЛИЗ КАЧЕСТВА УСТНОЙ РЕЧИ

Н.В. Карпов

Филиал национального исследовательского университета «Высшая школа экономики», Н. Новгород

nkarpov@hse.ru

Поступила в редакцию 15.05.2012

Рассмотрен и экспериментально исследован алгоритм автоматического анализа качества устной речи на основе критерия максимума скорости создания информации на выходе голосового тракта диктора. Синтезирован и экспериментально изучен новый алгоритм анализа качества речи с использованием кепстрального преобразования для параметризации сигнала.

Ключевые слова: качество речи, распознавание речи, авторегрессионная модель, кепстр.

Введение

В современных цифровых устройствах передачи и хранения данных используются методы, позволяющие значительно сжать речевой сигнал. В результате часть информации теряется, и, как следствие, качество звука может ухудшиться. В этой связи немалый интерес представляют методы анализа качества речи. Они позволяют оценить, насколько сильно искажен сигнал после прохождения в тракте связи по сравнению с исходным.

В соответствии с принятой терминологией качество речи – это величина, характеризующая субъективную оценку звучания речи. Таким образом, качество речи оценивается исключительно методом экспертных оценок. Определение качества речи с использованием автоматических алгоритмов позволит ускорить и удешевить этот процесс, что представляется интересной задачей. Целью данной статьи является синтез автоматического алгоритма для анализа качества слитной речи.

Для достижения этой цели в работе рассмотрены традиционные методы измерения качества речи и исследованы подходы для их автоматизации. Методы взяты из действующих стандартов [1, 2], которые устанавливают ряд норм качества передачи (воспроизведения) речи и методы их измерений.

Исследуется существующий метод автоматического анализа качества устной речи. На его основе синтезируется новый автоматический алгоритм анализа качества речи. Он экспериментально исследуется с использованием параметризации речевого сигнала кепстральными коэффициентами.

Теоретический анализ

Рассмотрим методы измерения показателей качества речевого сигнала для систем передачи данных согласно действующим стандартам. Основной метод оценки измерения качества речевого сигнала – это метод парных сравнений. Специальная бригада экспертов прослушивает некоторый набор слогов после прохождения по двум каналам связи и ставит оценки их звучания. При этом одинаковые оценки качества звука не допускаются.

В работе [3] приводится метод анализа качества речи на основе информационной теории восприятия речи. Как и в методе парных сравнений, в нем анализируемые сигналы сравниваются с отобранными образцами. Делать это предлагается при помощи величины информационного рассогласования (1) и обеляющего фильтра:

$$\rho_r(\mathbf{x}) = F^{-1} \times$$

$$\times \sum_{f=1}^{F} \frac{\left|1 - \sum_{m=1}^{p} a_r(m)\exp\frac{-j\pi mf}{F}\right|^2}{\left|1 - \sum_{m=1}^{p} a_x(m)\exp\frac{-j\pi mf}{F}\right|^2} - 1. \quad (1)$$

Здесь $\{a_x(m)\}$, $\{a_r(m)\}$ – векторы коэффициентов линейной авторегрессии (АР-коэффициентов) тестируемого сигнала $\mathbf{x}$ и эталона $\mathbf{x}_r^*$ класса r соответственно, оба одного порядка $p > 1$; f – дискретная частота, F – ее верхняя граница, или 1/2 частоты дискретизации речевого сигнала. Выражение в числителе (1) определяет квадрат амплитудно-частотной

характеристики *r*-го обеляющего фильтра, настроенного на *r*-й речевой образ $\mathbf{x}_r^*$, $r = \overline{1,R}$.

Как показано в работе [4], при гауссовом распределении сигналов и нормировке по дисперсии порождающего шума выражение для оптимального решающего правила в задаче R – альтернативной статистической классификации анализируемой выборки $\mathbf{X} = \{\mathbf{x}_m\}$ – сводится к виду

$$W_\nu(\mathbf{x}_m):\ \rho_r^{AR}(\mathbf{x}_m) = \sigma_r^2(\mathbf{x}_m)\Big|_{r=\nu} = \min. \quad (2)$$

Решение о классификации речевой единицы $\mathbf{x}_m$ принимается по критерию (2) минимума дисперсий откликов системы обеляющих фильтров (3) при $r = \overline{1,R}$:

$$\sigma_r^2(\mathbf{X}) = M^{-1}\sum_{m=1}^{M}\left[y_m^{(r)}(\mathbf{X})\right]^2,$$
$$y_m^{(r)}(\mathbf{X}) = \mathbf{A}_r^T\mathbf{x}_m, \quad (3)$$
$$\mathbf{A}_r = [1;-\mathbf{a}_r].$$

Здесь $\mathbf{a}_r$ – вектор коэффициентов авторегрессии, по которым находятся весовые коэффициенты $\mathbf{A}_r$ цифрового трансверсального фильтра с номером *r*, а $y_m^{(r)}(\mathbf{X})$ – сигнал на выходе того же фильтра при входном сигнале $\mathbf{x}_m$.

Для решения задачи качества устной речи величина информационного рассогласования рассчитывается для каждого речевого сегмента и предопределенного фонетического класса. При этом ряд сегментов считается не соотнесенным ни с одним из классов по заданному критерию превышения величины минимального информационного рассогласования наперед заданного порога

$$\min_r \rho_r(\mathbf{x}) > \rho_{порог}. \quad (4)$$

В работе [3] показано, что такой критерий качества эквивалентен критерию максимума скорости создания информации на выходе голосового тракта диктора. На основе этого теоретически обосновываются некоторые положения информационной теории качества речи, описанной там же.

Существует еще один способ измерения качества речи – это определение узнаваемости голоса. Этот метод относится к идентификации диктора по его голосу: эксперт должен узнать диктора по голосу из ограниченного набора лиц. Для автоматического распознавания диктора необходимо выделить такие параметры речи, которые будут всегда одинаковыми у одного человека и индивидуальными для разных людей при воспроизведении речи. Часто для этого используют кепстральное преобразование сигнала в методе LPCC (Linear Predication Cepstrum Coefficients) [4]. Это преобразование позволяет отделить характеристики фильтра $h(t)$ от исходного сигнала $x(t)$, которые присутствуют в речи в виде свертки $y(t) = x(t) * h(t)$. Для этого нужно выполнить следующие шаги:

- преобразование Фурье

$$x(t) * h(t) \to X(f)H(f); \quad (5)$$

- логарифмирование

$$X(f)H(f) \to \hat{X}(f) + \hat{H}(f);$$

- обратное преобразование Фурье

$$\hat{X}(f) + \hat{H}(f) \to \hat{x}(t) + \hat{h}(t).$$

Таким образом, общая формула для вещественного кепстра выглядит следующим образом:

$$C(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left|X(e^{i\omega})\right|e^{i\omega n}d\omega. \quad (6)$$

Метод предполагает предварительный этап обучения эксперта, производящего оценку. Можно сказать, что при обучении у него формируются некоторые образы, соответствующие каждому диктору.

В настоящее время благодаря ряду работ кепстральное преобразование используется в подавляющем большинстве систем распознавания и обработки речи. В этой связи разработка алгоритма анализа качества речи для сигнала, параметризованного кепстральными коэффициентами, представляется весьма актуальной задачей.

ГОСТ 16600-72 [2] дополнительно уточняет, что следует относить к ошибкам, а что нет. Например, перепутывание парных звонких и глухих согласных в слогах не является ошибкой. С точки зрения метода анализа качества речи это можно учитывать дублированием отдельных речевых единиц, вероятность перепутывания между которыми устанавливается близкой к единице.

Синтез алгоритма

На основе описанных идей постараемся синтезировать автоматический алгоритм анализа качества речи, применимый для широкого круга параметров речевого сигнала. При этом наша задача будет состоять в выборе такой шкалы или меры сравнения двух сигналов, на которой искаженный сигнал всегда будет иметь меньший уровень, чем неискаженный. Сравнивать сигналы будем, рассчитывая выбранную меру исходного сигнала и искаженного сигнала. Чтобы подтвердить эффективность предложенной меры, возьмем набор парных сигналов. При этом один сигнал из пары будет иметь заведомо худшее качество.

Для проведения экспериментального исследования метода анализа качества речи было записано нормальное и искаженное произнесение речи одного диктора. Для эксперимента выбрано сти-

Таблица 1

Дисперсии сигнала на выходе обеляющих фильтров

	1	2	3	4	5
А	3.385	6.093	2.454	3.015	4.366
В	3.114	**2.260**	6.577	3.089	6.788
Е	2.173	15.997	8.718	10.433	10.413
Ж	1.433	2.614	2.815	3.929	3.445
З	2.691	2.335	6.405	4.202	5.298
И	11.334	74.363	34.097	64.845	59.025
Н	4.069	4.246	2.178	5.988	3.503
М	3.243	18.076	6.051	10.022	21.263
О	1.517	3.958	4.237	2.543	4.970
Р	10.755	2.404	2.237	3.291	2.272
С	2.199	4.368	5.568	3.495	10.358
Ц	2.341	3.463	4.837	5.545	9.526
У	24.364	3.716	**2.166**	2.530	**2.092**
Ф	2.304	3.156	7.584	3.682	15.008
Х	3.646	4.465	13.837	3.875	19.405
Ч	2.722	25.070	10.645	13.018	15.418
Ш	1.661	7.995	6.477	**1.423**	8.561
Щ	4.221	34.345	12.715	22.836	22.458
Ы	**1.377**	4.604	3.537	5.336	3.498
Э	3.001	11.902	7.398	5.812	18.115

хотворение И.А. Бунина «Бушует полая вода». Текст был прочитан диктором сначала в привычном ритме, в обычных условиях, а затем после физических упражнений при нормальном дыхании. На слух эти два текста отличались не слишком сильно.

Протестируем записанные нами материалы с использованием метода, описанного в работе [3], и исследуем механизм его работы. При этом в качестве меры сравнения параметризованных речевых сегментов используем информационное рассогласование Кульбака–Лейблера и обеляющий фильтр.

После вычисления откликов M обеляющих фильтров на N входных сигналов получается матрица $M \times N$. В нашем случае это 20×5452. Рассмотрим фрагмент этой матрицы, содержащий пять первых столбцов (см. табл. 1). Каждый столбец в матрице – анализируемый речевой сегмент. Каждая строка – обеляющий фильтр, настроенный на соответствующую эталонную фонему. Число на пересечении – величина дисперсии сигнала после прохождения анализируемым сигналом обеляющего фильтра.

Минимальные значения величин информационных рассогласований выделены полужирным шрифтом. По критерию минимума информационного рассогласования первый сегмент речи будет отнесен в класс, соответствующий фонеме «Ы». Третий и пятый сегмент соотнесены с классом, который соответствует фонеме «У», четвертый – «Ш». Сегмент с номером два будет отбракован при выборе порогового значения меньше 2.2.

Для анализа результата классификации «ЫУШУ» сравним его с первым произнесенным словом «БУШУЕТ». При распознавании фонемы «Б» произошла ошибка, а следующие 3 фонемы распознаны правильно. Этот пример показывает, что хорошие результаты при таком подходе к распознаванию дают только фонемы, имеющие квазистационарные участки, а взрывные фонемы дают слабые результаты.

Для иллюстрации механизма распознавания рассмотрим два случая. В первом обеляющий фильтр согласован с фонемой, подаваемой на вход, и поэтому на выходе получаем маленькую дисперсию сигнала, которая меньше заданного порога $\rho_{порог}$ (рис. 1).

Из графика видно, что после обеляющего фильтра наблюдается снижение уровня спектральной мощности и выравнивание ее на всех частотах. Это подтверждает, что обеляющий фильтр был хорошо настроен на фонему.

Во втором случае обеляющий фильтр не согласован с фонемой, подаваемой на вход.

Из графика (рис. 2) после обеляющего фильтра видно, что спектр сигнала сильно неравномерный. Это говорит о том, что фильтр, настроенный на другой образ, не смог подавить фонему до белого шума.

В случае ухудшения качества сигнала рассогласование между входными сигналами и фильтрами, настроенными на качественный сигнал, будет увеличиваться. При этом даже минимальная дисперсия сигналов на выходе обеляющего фильтра (наилучшим образом настроенного) будут больше порога $\rho_{порог}$. Как

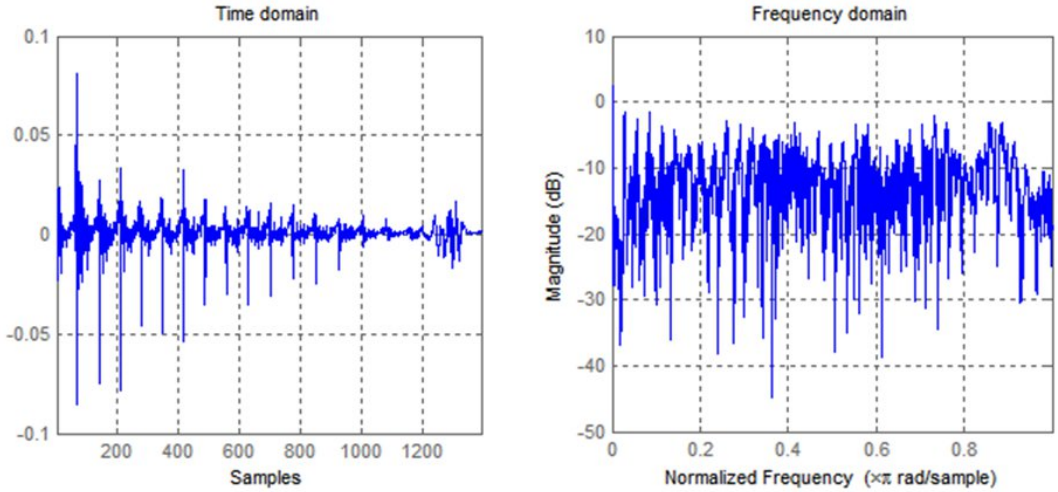

Рис. 1. Временная диаграмма и спектр гласного звука «А» после фильтра, настроенного на звук «А»

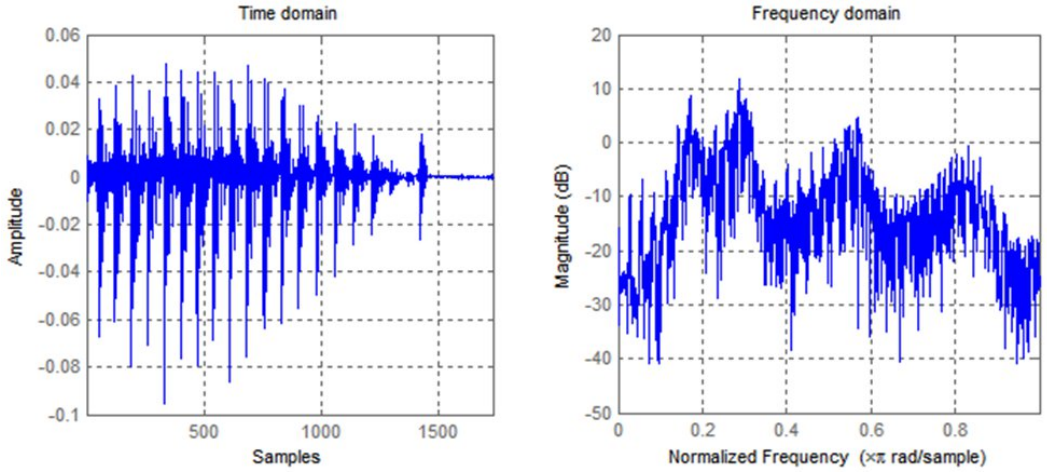

Рис. 2. Временная диаграмма и спектр фонемы «А» после фильтра, настроенного на другой звук («Ш»)

следствие, сегмент будет отбракован как некачественно проговоренный.

Очевидно, что количество отбракованных сегментов будет зависеть от выбора значения $\rho_{порог}$. Чем меньшим выбирается значение этого порога, тем больше сегментов попадут в категорию так называемых «плохо» проговоренных. Эта зависимость для двух сигналов отображена на рисунке 3.

По графику хорошо видно, что кривая, соответствующая речи после физических нагрузок, расположена выше. Это значит, что процент забракованных фонем после нагрузки при любом значении порога выше, чем в речи при нормальных условиях.

Алгоритм автоматического анализа качества речи с использованием параметризации речевого сигнала кепстральными коэффициентами предполагает ряд следующих шагов:

1. Формирование набора эталонных речевых образов, или рабочего словаря, длиной R

$$W_\nu(\mathbf{X}_L):\ L=\arg\left\{\min_j\left(\sum_{k=1}^{K}\rho_{jk}\right)\right\};\ \nu=\overline{1,R}. \quad (7)$$

В нашем эксперименте было записано R=20 речевых образов, соответствующих основным фонемам русского языка. Каждая фонема записывалась K=10 раз, после чего среди этого набора с помощью выбранного метода параметризации и евклидовой метрики находился центроид.

2. Запись речи диктора в цифровом виде и разделение ее на короткие сегменты квазистационарности, из которых формируется выборка для анализа

$$\mathbf{X}=\{\mathbf{x}_m\},\ m=\overline{1,M}. \quad (8)$$

Использовались сегменты без наложения друг на друга длительностью 10 мс. При частоте дискретизации $F_d=8000$ Гц длина каждого сегмента составила 80 дискретных отсчетов.

3. Кодирование каждого сегмента (8) выбранным способом параметризации

$$\mathbf{C}=\mathrm{F}(\mathbf{X})=\{\mathbf{c}_m\},\ m=\overline{1,M}. \quad (9)$$

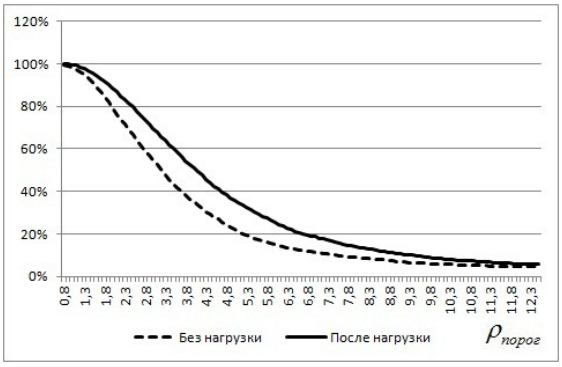

Рис. 3. Зависимость процента «плохо» проговоренных звуков от порогового значения $\rho_{порог}$

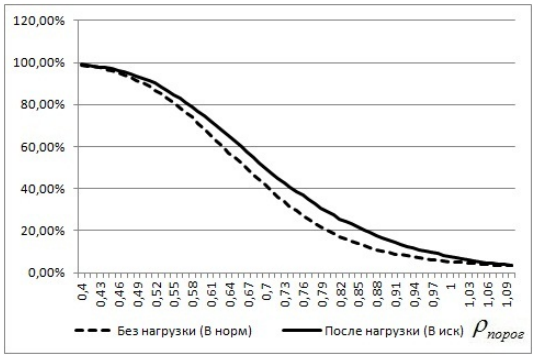

Рис. 4. Зависимость процента «плохо» проговоренных звуков от величины порога

Множество анализируемых сигналов $\mathbf{X} = \{\mathbf{x}_m\}$ и рабочий словарь объёмом $R > 1$ характеризуются авторегрессионной моделью (АР-модель) наблюдений, которая описывается следующей зависимостью:

$$x_r(n+1) = \sum_{i=1}^{p} a_r(i) x_r(n-i+1) + \varepsilon(n+1). \qquad (10)$$

Здесь $x_r(n+1)$ – значение $(n+1)$-го отсчета r-го речевого сигнала, $\{a_r(i)\} = \mathbf{c}_r^{AR}$ – вектор его АР-коэффициентов, p – порядок АР-модели, $\varepsilon_r(n+1)$ – процесс типа белого шума с нулевым значением математического ожидания и дисперсией σ_r^2, остающийся после фильтра с комплексным коэффициентом передачи

$$H_r(e^{j\omega}) = \frac{G}{1 - \sum_{i=1}^{p} a_r(i) e^{-j\omega i}}. \qquad (11)$$

Авторегрессионные коэффициенты вычисляются при помощи рекурсивной процедуры Берга–Левинсона. Далее они преобразуются в кепстральные коэффициенты

$$h_n = \begin{cases} a(n) + \sum_{k=1}^{n-1} \frac{k}{n} \hat{h}(k) a(n-k), & 0 < n < p, \\ \sum_{k=n-p}^{n-1} \frac{k}{n} \hat{h}(k) a(n-k), & n > p, \\ \ln G, & n = 0. \end{cases} \qquad (12)$$

Количество кепстральных коэффициентов N принято брать в диапазоне от 12 до 20 [5]:

$$\mathbf{c}_m^{Cepstr} = \{h_n\}, \quad n = \overline{1, N}. \qquad (13)$$

Таким образом было получено $M = 5452$ вектора параметров, каждый из которых характеризовал один из сегментов.

4. Классификация векторов в один из речевых образов (векторное квантование). Для вектора c_m определяем расстояния до всех речевых образов $\mathbf{c}_r$, $r = \overline{1, R}$, и находим минимум среди них:

$$W_\nu(\mathbf{X}): \ \rho_r^{Cepstr}(\mathbf{x}_m)\Big|_{r=\nu} = \min. \qquad (14)$$

Сегменты, параметризованные в кепстральные коэффициенты, соотносятся с речевыми образами при помощи евклидовой метрики

$$\rho_r^{Cepstr}(\mathbf{x}_m) = d_{Euclid}^2 = \\ = (\mathbf{c}_m^{Cepstr} - \mathbf{c}_r^{Cepstr})(\mathbf{c}_m^{Cepstr} - \mathbf{c}_r^{Cepstr})^T. \qquad (15)$$

5. Сравнение минимальных величин «расстояний» до всех имеющихся образов фонем с наперед заданным порогом и подсчет количества отбракованных сегментов для разных уровней

$$\min_r \rho_r > \rho_{порог}. \qquad (16)$$

Количество отбракованных сегментов B зависит от выбора значения $\rho_{порог}$. Образ или класс, имеющий минимальное расстояние, помечался как наиболее подходящий. Это минимальное евклидово расстояние сравнивалось с пороговым значением. В случае превышения порога сегмент, соответствующий такому вектору кепстральных коэффициентов, учитывался как недостаточно качественно произнесенный.

Экспериментальные исследования

Протестируем синтезированный алгоритм экспериментально. Для этого следуем синтезированному алгоритму, в котором используем те же звуковые файлы с качественно и не качественно проговоренным текстом «Бушует полая вода».

Варьируем величину $\rho_{порог}$ от 0 до 2. В диапазоне от 0.4 до 1.1 число отбракованных сегментов получилось различным. Зависимость относительного числа «плохо» проговорённых сегментов $B_{норм}$ и $B_{иск}$ от величины порога приведена на рис. 4.

Для кепстральных коэффициентов график получился немного другим, чем для обеляющего фильтра, но в целом тенденция сохранилась. Ко-

Рис. 5. Процент прироста числа отбракованных речевых сегментов от величины порога $\rho_{порог}$

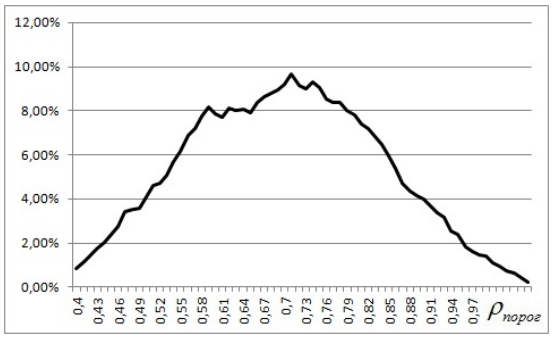

Рис. 6. Зависимость разницы $B_{иск} - B_{норм}$ от величины порога $\rho_{порог}$

личество отбракованных фонем после физической нагрузки $B_{иск}$ получается всегда больше, чем в нормальных условиях $B_{норм}$. Зависимость на рис. 4 можно интерпретировать как функцию распределения речевых единиц вокруг предопределенных эталонных речевых образов (7). В таком случае можно построить плотность функции распределения или зависимость процента прироста числа отбракованных речевых сегментов от величины порогового расстояния $\rho_{порог}$ (рис. 5).

Различие в качестве двух сигналов характеризуется расстоянием между двумя кривыми на рис. 4. Вычислим их разность $B_{иск} - B_{норм}$. Она будет показывать разницу количества «плохо» проговоренных сегментов в двух сигналах. Построим зависимость этой разности для каждой величины порога (см. рис. 6).

Количественно различия в качестве речи можно характеризовать величиной $B_{иск} - B_{норм}$, зафиксировав любое пороговое значение. Например, для $\rho^*_{порог}=0.7$ разность величин $B^*_{иск} - B^*_{норм} = 7.93\,\%$.

Заключение

В работе экспериментально исследованы два алгоритма автоматического анализа качества речи. Первый описан в работе [3] и разработан на основе критерия, введенного в информационной теории качества речи. Он использует авторегрессионные коэффициенты и обеляющий фильтр. На его основе синтезирован новый алгоритм автоматического анализа качества устной речи. Его отличительной особенностью является то, что определение величин расстояний (рассогласований) между двумя сигналами может производиться с использованием кепстральных коэффициентов и евклидовой метрики. В двух рассмотренных алгоритмах получаются в целом аналогичные и стабильные результаты.

Проведенное исследование показывает, что качество речи можно характеризовать количественно, используя синтезированный алгоритм, который позволяет вычислять относительное число отбракованных сегментов при фиксированной величине порога. Можно сделать предположение, что синтезированный алгоритм подходит для достаточно широкого класса методов параметризации речи и метрик.

Список литературы

1. ГОСТ Р 50840-95. Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. М: Изд-во стандартов, 1996. 234 с.

2. ГОСТ 16600-72. Передача речи по трактам радиотелефонной связи. Требования к разборчивости речи и методы артикуляционных измерений. М.: Стандартинформ, 2007. 76 с.

3. Савченко В.В. Информационная теория качества речи // Изв. вузов. Радиоэлектроника. 2011. Вып. 1. С. 22–32.

4. Furui Sadaoki. Digital speech processing, synthesis, and recognition. 2nd ed., rev. and expanded, 2000.

5. Карпов Н.В., Савченко В.В., Акатьев Д.Ю. Автоматическое распознавание элементарных речевых единиц методом обеляющего фильтра // Изв. вузов. Радиоэлектроника. 2007. Вып. 4. С. 11–19.

AUTOMATIC SPEECH QUALITY ANALYSIS

N.V. Karpov

The algorithm of automatic speech quality analysis is considered and experimentally studied on the basis of the maximum rate criterion of information production at the speaker vocal tract output. A new algorithm for automatic speech quality analysis is synthesized and experimentally tested using cepstral transformation for signal parameterization.

Keywords: speech quality, speech recognition, autoregressive model, cepstrum.