УДК 621.372 : 519.72

Н.В. Карпов

Национальный исследовательский университет - Высшая школа экономики в Нижнем Новгороде

Анализ качества речи в информационной теории восприятия речи.

Speech quality analysis using Informational Theory of Speech Perception.

Качество речи, обработка речи, кепстральные коэффициенты, обеляющий фильтр.

Speech quality, speech processing, MFCC, whitewash filter.

В статье проанализированы методы анализа качества речи, описанные в действующих ГОСТ и предложены более строгие и менее затратные на основе информационной теории восприятия речи. Экспериментально исследована эффективность и границы применимости трех современных методов параметризации речевых сигналов при использовании их с различными метриками.

In this article analyzed some methods of speech quality estimation based on State Standard and Informational Theory of Speech Perception. Experimentally examine effectiveness and boundaries of free methods for speech parameterization and using it with deferent metrics.

Введение. Анализ качества речи является важным для целого ряда прикладных задач. Основными документами, определяющими требования к качеству речи в трактах связи для передачи речи и в том числе в радиотелефонных траках связи являются Российский ГОСТ Р 50840-95 «Методы оценки качества, разборчивости и узнаваемости речи» и межгосударственный ГОСТ 16600-72 «Требования к разборчивости речи и методы артикуляционных измерений». В этих же стандартах описаны подходы к определению качественных характеристик речи все основанные на субъективной оценке параметров или другими словами на методе экспертных оценок. Недостатками его является, во-первых, дороговизна метода, а во-вторых, высокая степень зависимости от субъективных факторов самого эксперта. В этой связи представляется интересным использование более строгих и менее затратных подходов для решения задачи анализа качества речи.

Теоретический анализ. В терминах информационной теории восприятия речи (ИТВР) [1] опишем методы измерения показателей разборчивости и качества речевого сигнала, которые нужно контролировать для оценки качества речи, согласно ГОСТ Р 50840-95.

Первый метод – это оценка разборчивости речи методом артикуляционных измерений. Разборчивость оценивается с помощью

процента верно распознанных изолированных слогов или морфем. Каждый слог в ИТВР представляет собой речевой образ, причем один и тот же слог произносится и воспринимается на слух по-разному. Это так называемые речевые «метки» образа. На всем множестве звуков, доступных для восприятия человек способен различать разные речевые образы, и у каждого речевого образа есть некоторые границы, в которых он воспринимается как одинаковый.

Для оценки степени близости двух сигналов хорошие результаты дает применение информационное рассогласование Кульбака-Лейблера [2]. Она дает оценку величины взаимной информации двух случайных выборок по их закону распределения.

$$\rho_{12} = \sum_{i=1}^{m} f_1(i) \ln \frac{f_1(i)}{f_2(i)} \qquad (1)$$

где f_1, f_2 – законы распределения дискретных случайных величин $X_1(t), X_2(t)$. Вся их область значений разбивается на *m* равных отрезков (хотя разбиение может быть и не равномерным) и считается частота попадания значений $X(t)$ в *i*-ый эти отрезки.

На его основе разработан критерий минимума информационного рассогласования.

$$H_\nu(\mathbf{X}):\ \rho_r(\mathbf{X}) = \sigma_r^2(\mathbf{X})/\sigma_r^2 + \ln\sigma_r^2\,\big|_{r=\nu} = \min \qquad (2)$$

H_ν - ν-ая гипотеза о законе распределения дискретной случайной величины $\mathbf{X}$, $\sigma_r^2(\mathbf{X})$ – дисперсия на выходе обеляющего фильтра настроенного по сигналу с номером r из R эталонных, σ_r^2 – дисперсия возбуждающего белого Гауссова шума в авторегрессионной модели сигнала *r*. Решение здесь принимается в пользу ν-ой гипотезы при условии минимизации взвешенной с коэффициентом $1/\sigma_\nu^2$ и смещенной на $\ln\sigma_\nu^2$ величины выборочной дисперсии $\sigma_\nu^2(\mathbf{X})$ отклика на сигнал $\mathbf{X}$ обеляющего фильтра ν-го канала.

Этот критерий в стандартной формулировке об оценке статистических гипотез при допущении о Гаусовости закона распределения случайной величины в асимптотическом случае, как показано в работе [3], эквивалентен критерию максимального правдоподобия.

Критерий минимума информационного рассогласования хорошо физически реализуется с использованием метода обеляющего фильтра (ОФ). В работе [4] показано, что задача классификации в один из образов сводится к параллельной фильтрации обеляющими фильтрами, настроенными на эталонные сигналы. Решающее правило при этом

выражается формулой (3) как минимум дисперсии на выходе обеляющего фильтра с учетом нормировки по дисперсии порождающего шума.

$$H_\nu(\mathbf{X}):\ \rho_r(\mathbf{X}) = \sigma_r^2(\mathbf{X})\big|_{r=\nu} = \min \qquad (3)$$

Параметризация речевого сигнала при помощи коэффициентов линейного предсказания, позволяет выделять только полезную информацию для распознавания речи, содержащуюся в огибающей спектральной плотности мощности. Это немало важно для создания устройства анализа качества речи [5]. Существующие сегодня быстрые цифровые преобразования Фурье и метод Берга-Левинсона позволяют с минимальными вычислительными затратами реализовать этот критерий сравнения [4].

Другой способ описания огибающей спектральной плотности мощности – это вычисление мощности на выходе полосных фильтров с учетом того, что субъективно, разные частоты воспринимаются по-разному. Чем ниже частота, тем чувствительность выше. Это обусловлено биологическими особенностями человеческого механизма восприятия звука. Для описания этого введена специальная единица измерения – мел. Формула взаимосвязи частоты в герцах (f) и мелах (m) приведена ниже (4).

$$m = 1127{,}01048 \ln(1 + f/700) \quad (4)$$

Представленная формула определена в результате обработки большого массива данных субъективного восприятия высоты разных звуковых тонов.

Так как спектр звука имеет несколько периодических максимумов, поэтому можно применить еще раз преобразование Фурье и получить параметры, называемые кепстральными коэффициентами в мел-частотной шкале (MFCC) [11].

Второй метод оценки измерения показателей разборчивости и качества речевого сигнала речевого сигнала, - это разборчивость речи методом артикуляционных измерений по таблицам неполных слогов.

Этот метод предполагается использовать, при сильном зашумлении и низком уровне разборчивости передаваемого сигнала. Эксперту, производящему оценку разборчивости известен целиком слог кроме одной буквы, на которую тот либо начинается, либо заканчивается. Результатом является процент верно распознанных экспертом букв или фонем.

Для автоматизации этой работы можно использовать модифицированный метод обеляющего фильтра, описанный в работе [6], для задачи автоматического распознавания изолированных слов. Благодаря введенным дополнительным процедурам нормировки элементарных

речевых сигналов и ограничениям величины решающей статистики минимума информационного рассогласования по каждому отдельному сегменту слова (5)

$$\rho(\mathbf{X}) = \begin{cases} U, \text{ при } \sigma^2(\mathbf{X}) - 1 > U, \\ \sigma^2(\mathbf{X}) - 1, \text{ при } \sigma^2(\mathbf{X}) - 1 \le U \end{cases} \quad (5)$$

удается наилучшим образом учесть контекст, в котором произносится звук. За счет введения нормировки сигналов удается учесть существенное неравноценное значение фонем в формировании слитной речи.

Для всех описанных методов измерения показателей разборчивости, важным элементом является процедура предварительного формирования рабочего словаря. Эта задача подробно рассмотрена в работе [7]. В ней рассчитываются вероятность ошибки при классификации тестируемого сигнала, и анализируется, какие параметры словаря влияют на нее. Сделан вывод, что существенным образом на уровень ошибки влияет выбор наиболее типичной речевые метки в качестве эталона. Для строгого определения эталонной метки всего образа используется критерий минимума суммы информационных расстояний. Если проводить аналогию с физикой, то это будет центр масс подмножества точек.

$$r_{\min} = \arg(\min_r \sum_{i=1}^{N} \rho_{ri}) \quad (6)$$

В ГОСТ 16600-72 приводится методика артикуляционных измерений разборчивости на звукосочетаниях, словах и фразах. Задача автоматического определения степени разборчивости звукосочетаний в ИТВР решается при помощи метода фонетического декодирования слов описанного в [8]. В этом методе используется альтернативный подход, в отличие от применения дополнительной процедуры выравнивания слов по темпу речи на основе методов динамического программирования, как это делается в большинстве современных систем. Используется тонкая структура речевого сигнала, то есть распознавание происходит по «надежным» фонемам с учетом известной избыточности речевого сигнала. Применение этого метода позволяет значительно сократить суммарные вычислительные затраты на реализацию с почти 35 млн. операций до 4800 для сигнала длиной в 1 секунду.

Экспериментальным путем в работе [7] продемонстрировано малая зависимость вероятности ошибки алгоритма от настройки на конкретного диктора. Это, несомненно, тоже является важным фактором с точки зрения перспективы его применения.

ГОСТ 16600-72 дополнительно устанавливает ряд показатели качества речи: разборчивость звуков, процент правильно принятых звуков по

каждой таблице, среднее и среднее квадратичное отклонение разборчивости приема по каждой группе слов и каждому диктору.

В обоих действующих на сегодняшний день стандартах предполагается обучение бригады экспертов к новому тракту для передачи речи. Обучение проходит до тех пор, пока средняя частота правильного распознавания не стабилизируется, то есть не достигнет естественного для человека уровня.

ИТВР тоже предусматривает обучение системы без учителя. В статье [9] описаны базовые принципы информационной теории обучения речи. При этом обучение состоит в классификации каждого нового речевого сигнала, который в свою очередь добавляется в этот класс или речевой образ по критерию «достаточности степени близости» $\rho(X) \le \rho_0$. При этом он обогащает свой речевой образ, давая возможность в следующий раз более точно идентифицировать речевую метку. Вводится понятие информационного эталона как центр масс множества речевых меток. При дополнении образа очередной меткой, естественно сдвинется и информационный центр этого образа, что позволит при распознавании минимизировать ошибку.

В работе [10] дается теоретическое обоснование методу анализа качества речи на основе информационной теории восприятия речи. Предложен новый критерий качества устной речи: максимума скорости создания информации на выходе голосового тракта диктора.

ГОСТ 16600-72 уточняется требования к разборчивости речи, а именно, что следует относить к ошибкам, а что нет. Например, перепутывание парных звонких и глухих согласных в слогах не является ошибкой. С точки зрения описанного метода анализа качества речи [9] это учитывается дублированием отдельных речевых единиц, вероятность перепутывания между которыми устанавливается близкой к 1. То есть, пороговое значение информационного рассогласования выбирается по следующему правилу (7)

$$\rho_0 = \min_{r,\nu} \rho_{r\nu} \qquad (7)$$

Экспериментально была исследована эффективность применения различных способов кодирования и методов измерения расстояния между фонемами для задачи анализа качества речи. Эта задача ставилась как классическая задача классификации. Будем считать, что чем меньше вероятность ошибки при классификации, тем лучше применим данный метод к задаче анализа качества речи. Для этого, по аналогии с задачей проверки статистических гипотез, будем оценивать вероятности ошибок первого и второго рода при автоматической классификации различных фонем.

Экспериментальное исследование. Для постановки эксперимента было записано 43 фонемы русского языка, согласно Петербургской фонологической школе: [а э и о у ы п п' б б' м м' ф ф' в в' т т' д д' н н' с с' з з' р р' л л' ш ж щ ӂ ц ч й к к' г г' х х'], по 10 реализаций каждой фонемы. Формат цифровой записи был выбран wav, 8 кГц – частота дискретизации, моно сигнал, 16 бит. Программа обработки сигналов была написана с использованием среды математического моделирования Matlab.

Современные системы обработки речи используют приблизительно следующий набор параметров, описывающий фонемы:

1. Выделение информации из огибающей спектральной плотности мощности фонемы;
2. Нормализованная энергия фонемы;
3. Отношения энергии в нижней (0-3000 Гц) и верхней (3000-7500 Гц) полосе частот;
4. Средняя частота первой, второй третьей и четвертой форманты;
5. Величина амплитуды первой, второй третьей и четвертой форманты.

Экспериментально будем исследовать первый пункт из приведенного списка. Для этого было выбрано три способа кодирования фонем: коэффициенты линейного предсказания рассчитанные при помощи Берга-Левинсона, полосная мощность сигнала в мел-частотной шкале и кепстральные коэффициенты в мел-частотной шкале [11]. На основе этих коэффициентов мы оцениваем степень отличия одного сигнала от другого ρ_{jk}, используя следующие методы:

1. Информационное рассогласование Кульбака-Лейблера, реализованное при помощи обеляющего фильтра (ОФ) $\rho_{jk}^{O\Phi} = \sigma_k^2(\mathbf{X}_j)$;
2. Евклидова метрика (ЭМ) $\rho_{jk}^{Э} = \sum_{i=1}^{N}(x_i^{(j)} - x_i^{(k)})^2$;
3. Манхетеновская метрика (ММ) $\rho_{jk}^{M} = \sum_{i=1}^{N}| x_i^{(j)} - x_i^{(k)} |$;
4. Один минус коэффициент корреляции (Кор.)
$$\rho_{jk}^{kor} = 1 - \frac{X_j X'_k}{\sqrt{(X_j X'_j)(X_k X'_k)}};$$
5. Один минус коэффициент ковариации (Ков.)
$$\rho_{jk}^{kov} = 1 - \frac{\widehat{X}_j \widehat{X}'}{\sqrt{(\widehat{X}_j \widehat{X}'_j)(\widehat{X}_k \widehat{X}'_k)}},$$

где $\widehat{X}_k = X_k - \overline{X_k}$, а $\overline{X_k}$ есть среднее значение.

В соответствии с формулировкой задачи классификации, будем соотносить каждое наблюдение из анализируемого множества с одним из заранее определенных классов $W_i; i = 1 \ldots 43$. Зададим критерий, сопоставляющий каждой реализации выборки $X = x_j$ одну из имеющихся гипотез H_k как минимум расстояния до эталонной реализации каждого класса - ρ_{jk}, где k – номер класса.

$$H_k : \arg(\min_k(\rho_{jk})) \tag{8}$$

Под эталонной реализацией будем понимать реализацию, наиболее близкую к гипотетическому центру масс множества точек, относящихся к одному классу. Находить эталон будем как выборку, имеющую минимум суммы расстояний между реализациями.

$$m \leftarrow \arg(\min_m(\sum_{j=1}^{J} \rho_{jm})) \tag{9}$$

J – количество элементов одного класса. ρ_{jm} - расстояние между реализациями с номерами j и m.

Таблица верной классификации наблюдений при использовании различных методов сравнения и параметризации речевых сигналов представлена в таблице 1. Невысокие проценты правильной классификации фонем не являются показателем низкого качества распознавания. Они объясняются выбором большого числа исходных классов фонем для анализа, а также не использованием энергетических признаков. Для классификации не производился никакой отбор фонем с точки зрения квазистационарности, а рассматривались все типы для оценки применимости методов каждой фонеме, согласно Петербургской фонологической школе, многие из которых при восприятии речи могут быть перепутаны, без потери качества распознавания. Было рассмотрено три широко известных способа кодирования или параметризации речевого сигнала.

В первом способе параметризации мы рассчитываем величину спектральной плотности мощности сигнала в логарифмической шкале. Фильтруем его гребенкой треугольных фильтров (Рис. 1) в мел-частотной шкале.

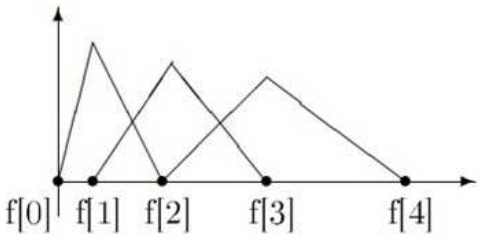


Рис. 1. Гребенка треугольных фильтров в мел-частотной шкале.

Далее, на выходе каждого фильтра, рассчитывается средняя мощность сигнала и получаются набор признаков, описывающий огибающую спектра сигнала. Их количество равно числу треугольных фильтров, расположенных в речевом диапазоне. Обычно это число выбирается между 13 и 20. Для нашего эксперимента мы взяли 20 фильтров. Наибольший процент правильной классификации для этих признаков дал сравнение при помощи коэффициента корреляции.

Во втором способе параметризации, для выделения кепстаральных коэффициентов используется логарифмическая спектральная плотность мощности после фильтрации треугольными оконными фильтрами полученными в первом методе. Далее применяется косинусное преобразование и получаются кепстральные коэффициенты. Наибольший процент при классификации получен при использовании манхетеновской метрики.

Третьим способом параметризации был расчет коэффициентов линейного предсказания (КЛП) рекуррентным методом Берга-Левинсона. При этом расстояние между двумя сигналами определяется по дисперсии на выходе обеляющего фильтра, настроенного на одни сигнал, на вход которого подается другой сигнал. При этом входной сигнал нормируется по дисперсии порождающего шума.

Таблица 1. Процент верной классификации фонем.

	ОФ	ЭМ	ММ	Кор.	Ков.
Полосная мощность сигнала	32,09	53,26	49,77	**61,40**	60,23
Кепстральные коэффициенты	21,63	53,95	**61,86**	59,77	58,60
КЛП	**62,56**	45,12	44,42	47,67	47,67

Итак, нам дана выборка $X=(X_1 \ldots X_N)$ из неизвестного класса и решается задача проверки гипотез H_k принадлежности i-ому классу, где i изменяется от 1 до 43, то есть соответствуют количеству возможных

фонем. Для понимания механизмов происходящего рассмотрим ошибки первого и второго рода.

Ошибкой первого рода, то есть «ложная тревога» будем называть случай, когда наблюдается реализация не из класса W_j, а из W_i и ошибочно принята гипотеза H_j, $i \neq j$. Ошибкой второго рода, то есть «пропуск цели» будем называть случай, когда наблюдается реализация из класса W_j, и ошибочно принята гипотеза H_i с другим номером $i \neq j$.

В таблице 2 приведены проценты ошибок первого и второго рода для трех методов, показавших лучшие результаты.

Таблица 2. Процент ошибок классификации I и II рода.

№		Кепстальн. коэфф.		Полосная мощность сигнала		КЛП и ОФ	
		I р.	II р.	I р.	II р.	I р.	II р.
1	А	0,00	0	0,00	0	0,00	10
2	Э	0,00	0	0,00	0	0,00	0
3	И	0,00	0	0,00	0	0,00	0
4	О	0,00	0	0,00	0	0,00	10
5	У	0,00	50	0,24	20	0,24	40
6	Ы	0,00	10	0,00	10	0,00	10
7	П	0,95	10	0,48	80	0,95	30
8	П'	1,43	30	0,95	50	0,71	70
9	Б	1,19	60	1,67	60	5,00	40
10	Б'	1,19	80	3,57	50	0,00	80
11	М	1,67	10	0,48	30	0,24	30
12	М'	1,67	40	0,48	30	0,71	80
13	Ф	1,43	70	0,95	90	3,57	60
14	Ф'	1,90	70	1,43	70	2,86	90
15	В	1,67	60	0,71	40	1,67	70
16	В'	0,48	30	1,43	50	0,24	50
17	Т	4,52	50	2,14	50	1,67	80
18	Т'	0,48	70	1,90	90	0,48	80
19	Д	0,71	70	0,24	60	0,71	50
20	Д'	0,48	40	2,62	30	0,48	20
21	Н	0,48	20	0,00	10	0,48	20
22	Н'	0,00	70	0,95	40	1,19	30
23	С	0,95	60	0,71	70	0,95	60

24	С'	1,19	40	0,95	40	1,67	10
25	З	4,05	10	5,24	10	1,19	20
26	З'	0,71	30	0,95	30	1,43	30
27	Р	0,71	30	0,24	10	0,71	0
28	Р'	0,00	20	0,24	20	0,48	0
29	Л	0,48	40	0,95	10	0,48	10
30	Л'	1,90	70	0,95	90	0,95	80
31	Ш	0,71	50	0,24	50	1,19	30
32	Ж	0,48	10	0,71	10	0,71	30
33	Щ	0,24	40	0,95	30	0,48	60
34	Ӂ	0,95	30	0,71	50	1,90	20
35	Ц	2,38	50	0,71	40	1,43	50
36	Ч	0,48	10	0,71	20	0,24	20
37	Й	0,24	30	1,19	10	0,24	50
38	К	0,95	10	1,19	10	0,71	20
39	К'	0,24	50	0,24	80	1,90	50
40	Г	0,95	50	0,71	40	0,24	70
41	Г'	0,95	50	1,19	50	0,24	40
42	Х	0,00	40	0,00	40	0,00	20
43	Х'	0,24	80	0,48	90	1,43	50

Результаты. Во всех трех методах вероятность ошибки 1 рода практически для всех вокализованных фонем равна нулю. Это говорит о том, что гласные хорошо классифицируются по огибающей спектральной плотности мощности и параметризация речевого сигнала во всех трех этих методах выбрана наилучшим образом.

Самые большие ошибки были получены у взрывных фонем «Б», «Т». Это объясняется тем, что они не являются квазистационарными сегментами речи и не могут иметь корректного описания ни коэффициентами линейного предсказания, ни рядом Фурье.

Достаточно большие ошибки первого и второго рода у шумовых фонем с небольшой амплитудой Х, Ш, С, Ф. Следует отметить, что ошибки II рода в шумовых фонемах Х, Ш, Ф при применении метода обеляющего фильтра меньше, чем в двух других, а ошибки I рода наоборот больше. Это является следствием применения процедуры нормировки по дисперсии порождающего шума.

Очевидно, что для верной классификации шумовых и взрывных фонем не достаточно только голосовых признаков. Необходимо их дополнить энергетическими или просодическими параметрами. Поэтому ошибки классификации шумовых и взрывных фонем можно

минимизировать, за счет введения дополнительной просодической информацией в фонетические признаки.

Таким образом, в статье предложен ряд подходов, на основе тех, что описаны в действующих ГОСТ, позволяющие получить более строгие и менее затратные оценки качества речи. Показано, что, в основном, они хорошо соотносится с информационной теорией восприятия речи. Также, экспериментально исследованы границы применения, наиболее известных методов параметризации огибающей спектральной плотности мощности. Выделен ряд фонем, для которых необходимо использовать дополнительную энергетическую или просодическую информацию для повышения качества анализа устной речи.

Список литературы

1. Савченко В.В.; Информационная теория восприятия речи, Известия вузов России. Радиоэлектроника, 2007, Вып. 6. С. 3–9.

2. Савченко В.В.; Автоматическая обработка и распознавание речи на основе принципа МИР и МОФ, Радиотехника и электроника. № 3, 2005,

3. Савченко В.В.; Обнаружение разладки случайного процесса по выборке на основе принципа минимума информационного рассогласования. Автометрия. №2, 2005, С. 68-74.

4. Савченко В.В., Акатьев Д.Ю., Карпов Н.В.; Автоматическое распознавание элементарных речевых единиц методом обеляющего фильтра. // Известия вузов России. Радиоэлектроника. Вып.4, 2007, С.11-19.

5. Furui, Sadaoki; Digital speech processing, synthesis, and recognition / Sadaoki Furui. - 2nd ed., rev. and expanded, 2000

6. Савченко В.В., Акатьев Д.Ю., Губочкин И.В.; Автоматическое распознавание изолированных слов методом обеляющего фильтра, Известия вузов России. Радиоэлектроника. Вып. 5, 2007.

7. Савченко В.В., Лукин П.Г.; Метод переопределенного словаря в задачах распознавания речевых сигналов, Радиотехника и электроника. Т.51. №2, 2006, С.202-206.

8. Савченко В.В.; Метод фонетического декодирования слов в задаче автоматического распознавания речи, Известия вузов России. Радиоэлектроника. Вып. 5 (сентябрь-октябрь), 2009, С. 41-49.

9. Савченко В.В.; Информационная теория обучения речи, Известия вузов России. Радиоэлектроника. Вып. 3 (май-июнь), 2009, С. 3-12

10. Савченко В.В.; Информационная теория качества речи, Известия вузов России. Радиоэлектроника. Вып. 1, 2011, С. 22-32.
11. Jacob Benesty, M. Mohan Sondhi, Yiteng Huan (Eds.); Springer Handbook of Speech Processing, Springer-Verlag Berlin Heidelberg, 2008, ISBN: 978-3-540-49125-5