Аналитическая оценка качества речи на выходе систем низкоскоростного кодирования при воздействии акустических помех

Кириллов С.Н., *доктор технических наук,*

Ромашкин Ю.Н., *кандидат технических наук,*

Картавенко Я.О.,

Дмитриев В.Т.

В статье рассматриваются известные способы аналитической оценки речи в системах низкоскоростного кодирования. Исследуются корреляционные связи этих оценок с результатами артикуляционных измерений разборчивости речи и узнаваемости голоса говорящего при передаче речи на фоне аддитивных помех. Находятся уравнения нелинейной регрессии, минимизирующие среднеквадратическую ошибку рассогласования расчётных и артикуляционных данных.

• низкоскоростное кодирование речи • разборчивость речи • узнаваемость голоса говорящего.

Well-known methods of speech quality analytical evaluation are considered for the systems with low-rate encoding. The correlation between these estimates and the results of articulatory measurements speech intelligibility and speaker recognition in the background of additive noise are examines. The non-linear regression equations minimizing the mean squared error the deviation of the calculated and articulator data are calculated.

• low-rate encoding • speech intelligibility • speaker recognition.

Введение

В настоящее время алгоритмы низкоскоростного кодирования речи находят широкое применение в различных системах передачи и приёма информации промышленного, военного и гражданского назначения. Требования и нормативные показатели по качеству передачи речевой информации постоянно ужесточаются в соответствии с расширением области применения систем. Современные радиосистемы с низкоскоростным кодированием обеспечивают слоговую разборчивость речи на 80–90%, однако зачастую за счёт потери узнаваемости [1].

Субъективное оценивание качества речи требует проведения большого количества артикуляционных испытаний, что, в свою очередь, приводит к значительным организационным и временным затратам. Кроме того, сильное влияние на результаты субъективного оценивания оказывают условия проведения испытаний, уровень окружающего шума, психоэмоциональное состояние аудиторов, степень тренированности бригады и другие факторы. Это приводит к невысокой повторяемости результатов и определённому разбросу полученных субъективных оценок.

Объективную оценку качества речи получают аналитическими методами или с помощью аппаратных средств. Это обеспечивает высокую повторяемость результатов, однако, в этом случае не в полной мере учитываются особенности слуховой системы человека. Кроме того, объективные методы оценки качества речи менее универсальны, чем субъективные.

Цель статьи — исследование функциональной связи ряда известных объективных оценок качества речи в низкоскоростных каналах связи с результатами артикуляционных измерений разборчивости речи и узнаваемости голоса говорящего при воздействии аддитивных помех различной интенсивности. Выбор на этой основе наиболее адекватных объективных оценок и их модификация для уменьшения ошибки рассогласования.

Объективные оценки качества речи

Известен ряд способов аналитической оценки качества приёма и передачи речи по различным каналам связи, включая и низкоскоростные. К наиболее широко применяемым можно отнести [2, 3]:

— расстояние Итакура-Саито (Itakura-Saito Distance — ISD):

$$\mathrm{ISD} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{|X(f_n)|}{|Y(f_n)|} - \log\frac{|X(f_n)|}{|Y(f_n)|} - 1\right), \quad (1)$$

где $X(f_n)$ и $Y(f_n)$ — средние спектры входного и выходного сигналов;

— искажение спектра барков (Bark Spectral Distortion — BSD):

$$\mathrm{BSD} = \frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N_k}\left(|X(f_n,k)| - |Y(f_n,k)|\right)^2, \quad (2)$$

где $X(f_n,k)$ и $Y(f_n,k)$ — средние спектры сигналов в *k*-й критической полосе частот, N_k — количество спектральных отсчётов в *k*-й критической полосе, *K* — общее количество критических полос;

— модифицированное искажение спектра барков (Modified Bark Spectral Distortion — MBSD):

$$MBSD = \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}B(k)[X(f_n,k) - Y(f_n,k)]^2, \quad (3)$$

где *B(k)* — показатель ощущения искажений в *k*-й полосе (равен 0, когда искажения в полосе не воспринимаются на слух, и равен 1 в противном случае);

— корреляция средних спектра (Excitation Spectral Correlation — ESC):

$$ESC = \frac{\left(\sum_{n=1}^{N}|X(f_n)||Y(f_n)|\right)^2}{\sum_{n=1}^{N}|X(f_n)|^2\sum_{n=1}^{N}|Y(f_n)|^2}; \quad (4)$$

— фонетическая функция (функция ощущения спектральной динамики — ФОСД) предложенная А.А. Пироговым [4–6]:

$$ФОСД = \frac{1}{MN}\sum_{m=2}^{M}\sum_{n=1}^{N}\lg\left(\frac{\left|Y\left(f_n,m\right)\right|}{\left|Y\left(f_n,m-1\right)\right|}\right) - \sum_{m=2}^{M}\sum_{n=1}^{N}\lg\left(\frac{\left|X\left(f_n,m\right)\right|}{\left|X\left(f_n,m-1\right)\right|}\right). \quad (5)$$

Условия проведения экспериментов

Для предварительного тестирования рассматриваемых объективных оценок были сформированы аддитивные смеси речевого сигнала (слоговые и фразовые артикуляционные таблицы из [7]) с флуктуационными помехами, принадлежащими к следующим четырём классам:

— широкополосная стационарная (ШП СТ), в качестве которой использовался белый гауссовский шум (БГШ);

— низкочастотная стационарная (НЧ СТ) в виде БГШ на выходе ФНЧ-фильтра, АЧХ которого имеет частоту среза 2 кГц и наклон 9 дБ/октаву в сторону высоких частот;

— низкочастотная нестационарная (НЧ НСТ), представляющая собой реализацию шума транспортного потока;

— широкополосная нестационарная (ШП НСТ) в виде записи музыки.

Речевой сигнал и помехи в полосе частот (0,1–8,0) кГц преобразовывались на этом этапе в цифровой вид без сжатия (ИКМ кодирование) с частотой дискретизации 22050 кГц и разрядностью квантования 16 бит. Тестовые смеси формировались так, чтобы обеспечить значения отношения сигнал/шум (ОСШ) во входном сигнале в интервале от 12 до — 15 дБ с шагом 3 дБ.

С использованием этих записей далее в соответствии с [7] проводились артикуляционные измерения слоговой разборчивости речи (*S*,%) и узнаваемости голоса диктора (*U*, баллы). По полученным данным для каждого класса помехи вычислялся выборочный коэффициент корреляции между аналитическими расчётами (A_i), выполненными по формулам (1–5), и результатами артикуляционных измерений слоговой разборчивости речи (узнаваемости голоса говорящего):

$$\rho = \frac{\sum_{i=1}^{L}(S_i-\overline{S})(A_i-\overline{A})}{\sqrt{\sum_{i=1}^{L}(S_i-\overline{S})^2(A_i-\overline{A})^2}},$$

где *L = 10* — количество единичных измерений.

Результаты таких вычислений представлены ниже в табл. 1.

Таблица 1

Помеха		*ISD*	*BSD*	*MBSD*	*ESC*	*ФОСД*
ШП СТ	*S*	0,89	0,86	0,85	0,88	0,91
	U	0,86	0,85	0,78	0,87	0,90
НЧ СТ	*S*	0,06	0,90	0,9	0,89	0,93
	U	0,03	0,86	0,91	0,91	0,91
НЧ НСТ	*S*	0,29	0,33	0,48	0,93	0,91
	U	0,33	0,45	0,80	0,93	0,86
ШП НСТ	*S*	0,93	0,82	0,75	0,93	0,87
	U	0,90	0,72	0,87	0,90	0,88

Согласно полученным данным, среди рассматриваемых способов объективной оценки качества речи в целом можно выделить два: ESC для случаев воздействия нестационарных помех и ФОСД при наличии стационарной помехи.

Для этих двух способов дополнительно исследовалась возможность повышения корреляции с результатами артикуляционных измерений путём модификации, учитывающей различную чувствительность слуха по частоте [8]. Вводя коэффициент значимости критической полосы спектра $\beta(k)$ как долю энергии речи в *k*-критической полосе и используя аппроксимацию среднего спектра формулой [9]:

$$X(f_n)=\frac{4\sigma^2 f_n}{f_0\left(1+\left(f_n/f_0\right)^2\right)^2},$$

где σ^2 — множитель, задающий громкость речи, f_0 — среднее значение частоты основного тона, получим значения $\beta(k)$, которые приведены в табл. 2.

Таблица 2

k	$\beta(k)$	k	$\beta(k)$	k	$\beta(k)$	k	$\beta(k)$	k	$\beta(k)$
1	0.005	5	0.093	9	0.066	13	0.041	17	0.026
2	0.037	6	0.091	10	0.062	14	0.036	18	0.024
3	0.070	7	0.089	11	0.052	15	0.032	19	0.022
4	0.084	8	0.078	12	0.046	16	0.028	20	0.018

Тогда выражения для модифицированных оценок примут вид:

$$MESC=\frac{1}{K}\sum_{k=1}^{K}\frac{\left(\sum_{n=1}^{N_k}\beta_k\left|X(f_n,k)\right|\cdot\left|Y(f_n,k)\right|\right)^2}{\sum_{n=1}^{N_k}\beta_k\left|X(f_n,k)\right|^2\sum_{n=1}^{N_k}\beta_k\left|Y(f_n,k)\right|^2}, \quad (6)$$

$$МФОСД=\sum_{m=2}^{M}\sum_{k=1}^{K}\beta_k\sum_{n=1}^{N_k}\left(\lg\left(\frac{\left|Y(f_n,k,m)\right|}{\left|Y(f_n,k,m-1)\right|}\right)-\lg\left(\frac{\left|X(f_n,k,m)\right|}{\left|X(f_n,k,m-1)\right|}\right)\right). \quad (7)$$

Очевидно, что, когда спектры сигналов на входе и выходе совпадают, MESC=1 и МФОСД=0. В противном случае значения MESC всегда положительны и, как правило, не меньше 0,25; а МФОСД может принимать как положительные, так и отрицательнык значения в неопределённых пределах. Введение такой модификации позволило для обоих способов в среднем на 0,04 увеличить значение выборочного коэффициента корреляции между объективными и субъективными оценками как по критерию разборчивости речи, так и по узнаваемости голоса.

Следующая часть экспериментов была связана с применением формул (6) и (7) к оцениванию качества речи на выходе низкоскоростных кодеков. Структурная схема таких экспериментов показана на рис.1.

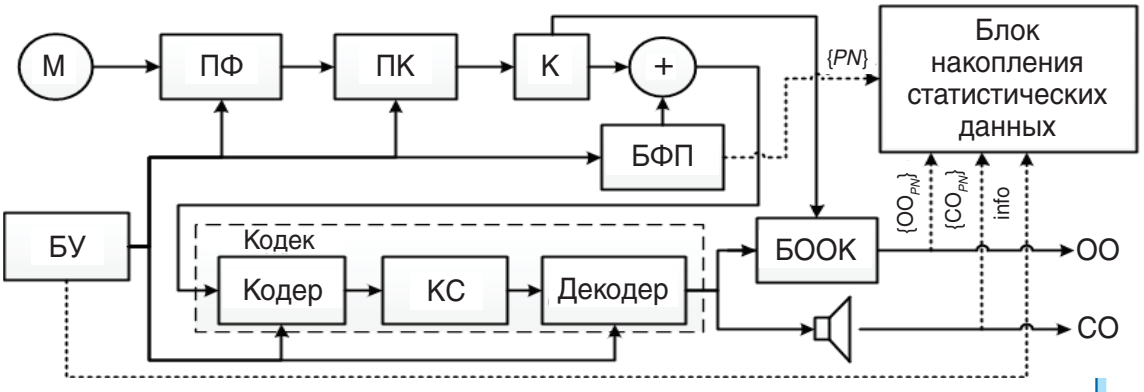

Рис.1. Структурная схема экспериментов

В качестве речевого материала использовались слоговые артикуляционные таблицы и тестовые фразы, приведённые в ГОСТ Р 50840–95 и начитанные 10 дикторами. Запись речи осуществлялась в помещении кабинетного типа объёмом 50 м^3 и временем реверберации 0,35 с при наличии слабого фонового шума. Микрофон М устанавливался на расстоянии 0,5 м перед диктором. Регистрируемый им речевой сигнал поступал на полосовой фильтр ПФ с полосой пропускания (0,3–3,4), (0,1–7,0) или (0,1–8,0) кГц и далее оцифровывал на ПК с разрядностями квантования 16 бит и частотой дискретизации 8000, 16000 или 22050 Гц соответственно.

Блок формирования помехи БФП осуществлял считывание с ПК реализации заданной акустической помехи, которая затем суммировалась с речью, обеспечивая требуемое значение ОСШ. В блоках КОДЕР и ДЕКОДЕР осуществлялось кодирование и декодирование тестовой смеси в соответствии с стандартизованными алгоритмами, которые были разделены на три группы:

1) кодеки со скоростью кодирования (1–16) кбит/с и полосой пропускания сигнала (0,3…3,4) кГц: LBRAMR, MMBE, ICELP, G723.1, G729a, GSM, G726, G728i;

2) AMR кодеки со скоростью кодирования (6,6–23,85) кбит/с и полосой пропускания сигнала (0,1…7,0) кГц;

3) кодеки со скоростью кодирования (32–64) кбит/с и полосой пропускания сигнала (0,1…8,0) кГц: MPEG 1.2.8 и Vorbis OGG.

Декодированный сигнал поступал на динамик для прослушивания аудитором и получения субъективной оценки (СО), а также на вход блока БООК, где вычислялась объективная оценка (ОО) качества речи по формулам (6) и (7). Блок БУ осуществлял выбор режимов работы ряда других блоков.

Далее находилась функциональная зависимость между расчётными (объективными) оценками качества речи и результатами артикуляционных (субъективных) измерений слоговой разборчивости речи и узнаваемости голоса говорящего. Поиск такой зависимости осуществлялся в классе уравнений полиномиальной регрессии P-го порядка:

$$CO = \sum_{p=0}^{P} \alpha_p \cdot OO^P . \qquad (8)$$

Коэффициенты α_p регрессии вычислялись по методу наименьших квадратов, обеспечивая минимум среднеквадратической ошибки для каждого из анализируемых кодеков. При этом ориентировались на достижение среднеквадратической погрешности не более 7,5% для слоговой разборчивости речи и 0,5 балла для узнаваемости голоса говорящего.

Результаты экспериментов

Объективные оценки, полученные при воздействии определённой помехи, сначала усреднялись по набору кодеков каждой группы, а затем для вычисленных средних значений строилась соответствующая кривая регрессии для слоговой разборчивости речи и узнаваемости голоса говорящего.

Ниже приведён ряд соответствующих примеров:

1) кодеки первой группы, помеха ШП НСТ:

$$S \approx 155.3MESC^2 - 51.8MESC + 9.5$$

$$U \approx 4MESC^2 - 0.44MESC + 0.85 .$$

Так, при *MESC* = 0,3 имеем *S* ≈ 7% и *U* ≈ 1,3, а при *MESC* = 0,9 *S* ≈ 90% и *U* ≈ 4.5.

2) кодеки второй группы, помеха ШП СТ:

$$S \approx -0.013МФОСД^2 - 0.4МФОСД + 98,$$
$$U \approx -0.005МФОСД^2 - 0.05МФОСД + 4{,}8.$$

При МФОСД = 0 имеем *S* ≈ 98% и *U* ≈ 4,8.

3) кодеки третьей группы, помеха НЧ НСТ:

$$S \approx -17.0MESC^2 + 95.1MESC - 15.0,$$
$$U \approx -5.4MESC^2 - 0.37MMESC + 0{,}5.$$

При *MESC* = 0,3 имеем *S* ≈ 15% и *U* ≈ 0.9, а при *MESC* = 0,9 *S* ≈ 84% и *U* ≈ 4.5.

Более точные результаты, удовлетворяющие указанным выше требованиям, вследствие различных принципов кодирования речи удалось получить при использовании для каждого кодека и вида помехи индивидуального уравнения регрессии. При этом заданная точность обеспечивалась при использовании уравнений регрессии второго (редко третьего) порядка. Значения максимальной абсолютной ошибки, полученной в этих экспериментах для некоторых кодеков, приведены в таблицах 3 и 4.

Таблица 3

Оценка MESC

	LBRAMR, 2 кбит/с		*G729a, 8 кбит/с*		*G 726, 16 кбит/с*	
	ΔS, %	*ΔU, балл*	*ΔS, %*	*ΔU, балл*	*ΔS, %*	*ΔU, балл*
ШП СТ	6,8	0,4	4,2	0,4	2,1	0,4
НЧ СТ	5,4	0,4	3,8	0,3	3,3	0,3
НЧ НСТ	3,9	0,3	0,9	0,3	1,5	0,3
ШП НСТ	3,2	0,3	2,1	0,3	2,1	0,1

Таблица 4

Оценка МФОСД

	ICELP, 4,8 кбит/с		*GSM, 13 кбит/с*		*MPEG, 56 кбит/с*	
	ΔS, %	*ΔU, балл*	*ΔS, %*	*ΔU, балл*	*ΔS, %*	*ΔU, балл*
ШП СТ	3,8	0,3	0,4	0,3	2,3	0,3
НЧ СТ	4,2	0,3	0,9	0,4	5,4	0,3
ШП НСТ	5,1	0,5	0,3	0,1	2,1	0,4
НЧ НСТ	4,7	0,4	2,1	0,3	5,3	0,4

Заключение

Рассмотрен ряд известных способов аналитической оценки качества речи, принятой на фоне аддитивной помехи. Экспериментально установлено, что наиболее адекватными являются оценки MESC (вычисление коэффициента корреляции между средними спектрами сигналов на входе и выходе системы) и МФОСД (вычисление функции ощущения спектральной динамики). Предложены модификации обоих способов, учитывающие зависимость чувствительности слуха человека по частоте. Показано, что оценку МФОСД целесообразно применять в случаях воздействия стационарных помех, а MESC — нестационарных.

Для ряда стандартизованных низкоскоростных кодеков со скоростью кодирования от 1 до 64 кбит/с получены результаты артикуляционных измерений слоговой разборчивости речи и узнаваемости голоса говорящего для четырёх классов аддитивных акустических помех. С использованием этих результатов найдены выборочные уравнения регрессии второго порядка, обеспечивающие точность аналитических оценок для каждого кодека, сравнимую с артикуляционными измерениями.

Список литературы

1. *Цыбулин М.К., Бочаров М.О.* Анализ методов оценки качества передачи речевой информации по каналам связи различной структуры. Электросвязь, 2008. № 11. С. 46–48.

2. *Wang S., Skey A., Gersho A.* An objective measure for predicting subjective quality of speech coders // IEEE Journal on Selected Areas in Communications, 1992. V. 10(5), P. 74–77.

3. *Ozer H., Avcibas I., Sankur B., Memon N.* Steganalysis of audio based on audio quality metrics // SPIE Electronic Imaging Conf. on Security and Watermarking of Multimedia Contents, 2003. Pp. 55–66.

4. *Пирогов А.А.* Синтетическая телефония. М.: Связьиздат, 1963.

5. *Пирогов А.А.* Вокодерная телефония. М.: Связь, 1974.

6. *Соболев В.Н.* Информационные технологии в синтетической акустике. М.: ИРИАС, 2007.

7. ГОСТ Р 50840–95. Передача речи по трактам связи. М.: Госстандарт России, 1995.

8. *Шелухин О.И., Лукьянцев Н.Ф.* Цифровая обработка и передача речи. М.: Радио и связь, 2000.

Сведения об авторах

Кириллов С.Н. —
доктор технических наук, Рязанский государственный радиотехнический университет

Ромашкин Ю.Н. —
кандидат технических наук, окончил Московский инженерно-физический институт, факультет «Автоматика и электроника». Область научных интересов: цифровая обработка речевых сигналов, фильтрация речи на фоне помех, автоматическое распознавание речи и языка, идентификация говорящего по голосу, низкоскоростное кодирование речи, оценка качества трактов речевой связи. E-mail: romayn@yandex.ru

Картавенко Я.О.
Рязанский государственный радиотехнический университет

Дмитриев Т.В.
Рязанский государственный радиотехнический университет