УДК 621.391:519.72

И. Н. ПРЕСНЯКОВ, д-р техн. наук, А. В. ОМЕЛЬЧЕНКО, канд. техн. наук, А. И. ПРЕСНЯКОВ

ОЦЕНКА КАЧЕСТВА РЕЧИ В ЦИФРОВЫХ СИСТЕМАХ ПЕРЕДАЧИ

Исторически первым критерием, по которому стали оценивать качество передачи речи, была громкость [1]. Однако уже в начале 30-х годов прошлого века проявился недостаток этого критерия, так как зачастую более громкая речь характеризуется меньшей разборчивостью, чем тихая. Для цифровых систем передачи речи такой показатель вообще не является содержательным.

При разработке систем связи качество переданной речи обычно характеризуют степенью разборчивости и степенью натуральности, которые определяются методом экспертных оценок. Однако для многих приложений, в частности в задачах синтеза систем связи, требуется иметь численные значения показателей разборчивости и натуральности речи, которые могли бы определяться на ЭВМ по заданным алгоритмам без привлечения бригады аудиторов. К сожалению, в настоящее время в известной авторам литературе такие показатели не описаны.

Данная статья посвящена обоснованию показателей разборчивости и натуральности речи применительно к задачам оптимизации цифровых систем передачи.

1. Показатели разборчивости речи

Созданная Коллардом в 20-х годах теория разборчивости речи была развита в ряде работ [1, 2], авторы которых предложили несколько отличных друг от друга методов расчета разборчивости.

В электросвязи под разборчивостью речи понимается ее свойство сохранять смысловую информацию. В качестве меры разборчивости обычно используется отношение числа правильно воспринятых слушателем элементов речи (звуков, слогов, букв, фраз) к числу переданных. В соответствии с типом используемых элементов различают разборчивость звуков, разборчивость слогов, разборчивость слов и разборчивость фраз. Для измерения указанных видов разборчивости наибольшее применение получил артикуляционный метод, предполагающий использование специально укомплектованных и обученных артикуляционных бригад.

Особое место в теории разборчивости занимает мера разборчивости формант, которая характеризует долю формант, воспринятых слушателем, от их общего числа, содержащегося в выборке исходной речи. В отличие от перечисленных ранее мер разборчивости она непосредственно не измеряется (в рамках классических подходов), однако является единственной из всех перечисленных мер разборчивости, которая может быть аналитически рассчитана.

Для задач синтеза и оптимизации систем обработки речевых сигналов требуется определить показатель разборчивости речи, который бы измерялся автоматически без участия человека. В качестве такого показателя в настоящей работе предлагается использовать модифицированный вариант классической меры разборчивости формант [1]. Для измерения этого показателя используются алгоритмы оценивания параметров формант, разработанные в рамках модели линейного предсказания речевых сигналов [3].

Покажем связь классической меры разборчивости формант и ее модифицированного варианта, предлагаемого к использованию в задачах оптимизации телекоммуникационных систем.

В работе [1] определен показатель формантной разборчивости речи

$$A = \sum_{k=1}^{n} A_k P_k \qquad (1)$$

где $A_k, k = \overline{1,n}$ – средние вероятности появления формант в k-ой частотной полосе, взвешенные так, что $\sum_{k=1}^{n} A_k = 1$; P_k – коэффициент восприятия, трактуемый как доля формант в к-ой частотной полосе, которая в заданных условиях будет воспринята слушателем с уровнем выше порогового.

В той же работе [1] обоснована методика измерения разборчивости речи. При этом обычно частотный диапазон речи разбивается на равноартикуляционные полосы, соответствующие одинаковым приращениям разборчивости формант. В данном случае все коэффициенты $A_к$ принимают одинаковые значения, и показатель разборчивости определяется следующей формулой (1)

$$A = \frac{1}{n} \sum_{k=1}^{n} P_k , \qquad (2)$$

где n – число равноартикуляционных полос, которое для диапазона речи 100-10000 Гц принято считать равным 20.

Значения коэффициентов восприятия в (1,2) являются функцией известного вида

$$P_k = f(E_k), \tag{3}$$

аргумент которой называется эффективным уровнем ощущения формант и определяется выражением

$$E_k = 10(\ell g S_k - \ell g S_{0k}), \tag{4}$$

где $S_к$ – уровень интенсивности формант в *к*-ой полосе частот; $S_{0к}$ – пороговый уровень интенсивности формант в *к*-ой полосе частот. Значение $S_{0к}$ определяется уровнем порога слышимости, затуханием всего тракта от микрофона до уха слушателя, маскирующим влиянием помех, и некоторыми другими факторами.

Таким образом, выражение для коэффициента восприятия может быть представлено в следующем виде:

$$P_k = f(10\ell g\alpha_k), \tag{5}$$

где $\alpha_k = \frac{S_k}{S_{0k}}$ – относительный уровень формант в к-ой полосе частот.

В задачах передачи речевых сообщений без потери общности, можно считать, что максимум разборчивости достигается при $S_к=S_{0к}$, $k = \overline{1,n}$, то есть функция $f(\alpha)$ принимает максимальное значение при нулевом аргументе. В этом случае величины $S_{0к}$ и $S_к$ имеют смысл формантного уровня в k-ой полосе для исходного речевого сигнала и для сигнала на выходе системы передачи соответственно.

Для задач синтеза и анализа систем связи наибольший интерес представляет случай высоких уровней разборчивости ($A \geq 0.95$). В этом случае возможна аппроксимация функции (5) усеченным рядом Тэйлора по переменной α_k

$$P_k \approx f(0) + a\cdot(\alpha_k - 1)^2 = f(0) + a\cdot(\frac{S_k}{S_{0k}} - 1)^2 = f(0) + a\cdot\frac{(S_k - S_{0k})^2}{S_{0k}^2}, \tag{6}$$

где постоянная

$$a = \frac{1}{\ln^2 10} 100\cdot f''(X)\Big|_{X=0} < 0\cdot \tag{7}$$

Подставляя (6) в (2) получим следующие приближение для показателя разборчивости

$$A \approx f(0) + a\cdot\frac{1}{n}\sum_{k=1}^{n}\left[\frac{S_k - S_{0k}}{S_{0k}}\right]^2, \tag{8}$$

которое справедливо при высоких уровнях разборчивости. Отсюда следует, что в оговоренных условиях вместо показателя разборчивости А может использоваться величина

$$\alpha = \frac{1}{n}\sum_{k=1}^{n}\left[\frac{S_k - S_{0k}}{S_{0k}}\right]^2, \tag{9}$$

которая характеризует погрешность воспроизведения формант.

Исходя из анализа особенностей восприятия речи и опыта существующих подходов к оцениванию ее разборчивости, сформулируем основные требования к показателю разборчивости речи: он должен учитывать точность воспроизведения не всех спектральных составляющих речи, а лишь ее формантных составляющих; он должен удовлетворять принципу аддитивности к элементарным частотным полосам; он в одинаковой мере должен учитывать как спектральные составляющие формант большой интенсивности, так и спектральные составляющие формант малой интенсивности (т.е. не должно быть маскирующего влияния интенсивных формант на форманты малой интенсивности); показатель разборчивости должен описываться простым аналитическим выражением.

Указанным требованиям удовлетворяет предложенный показатель (9) в виде средней погрешности представления формант по всему ансамблю речевых сигналов. С уменьшением значения α достигается более высокая разборчивость речи, переданной с использованием системы связи. В даль-

нейшем используем величину α в качестве показателя разборчивости при оптимизации систем передачи речевых сообщений.

Модифицируем показатель разборчивости (9), приведя его к виду, удобному для оптимизации систем передачи речевых сообщений.

Известно [3-5], что на коротких отрезках длительностью $T_C \approx 20$ мс речевой сигнал с допустимой погрешностью можно рассматривать как стационарный процесс. Представим реализации сигнала $x(t)$ длительностью T_n на выходе системы передачи речи в виде совокупности $p = [T_n / T_c]$, следующих друг за другом коротких сегментов сигнала

$$x_r(t) = x(t - (r-1)\cdot T_c), \qquad t \in [0, T_c);\ r = \overline{1, p}. \tag{10}$$

Таким же образом разобьем и исходный речевой сигнал на входе системы

$$x_{or}(t) = x_o(t - (r-1)\cdot T_c), \qquad t \in [0, T_c);\ r = \overline{1, p}. \tag{11}$$

Погрешность формантного представления каждого из таких сегментов речи охарактеризуем следующим аналогом показателя (9):

$$\tilde{\alpha}_r = \frac{1}{J}\sum_{j=1}^{J} \frac{\left[S_r(\hat{f}_j) - S_{0r}(\hat{f}_j)\right]^2}{S_{0r}^2(\hat{f}_j)}, \tag{12}$$

где J – число формант речи, учитываемых на рассматриваемом сегменте; $\{\hat{f}_j, j = \overline{1, J}\}$ – оценки центральных частот формант речи, найденные по соответствующим сегментам исходного речевого сигнала с использованием одного из известных алгоритмов [3]; $S_{or}(f), S_r(f)$ – значения авторегрессионного спектра r–го сегмента для исходного сигнала и сигнала на выходе системы передачи соответственно.

Вычисление значений авторегрессионных спектров в (12) осуществляется в соответствии с выражением [3–5]

$$S_r(f) = \frac{\Delta t \cdot \sigma_r^2}{\left|1 + \sum_{m=1}^{M} a_r(m)\cdot \exp(-j2\pi f \cdot \Delta t \cdot m)\right|^2}, \tag{13}$$

где σ_r^2 – оценка дисперсии сигнала, возбуждающего голосовой тракт на r-ом сегменте; $\{a_r(m), m = \overline{1, M}\}$ – совокупность оценок коэффициентов авторегрессии для r-ого сегмента.

В качестве показателя разборчивости всей совокупности сегментов речевого сигнала используем величину

$$\hat{\alpha} = \frac{1}{p}\sum_{r=1}^{p} \tilde{\alpha}_r, \tag{14}$$

где значения погрешности представления отдельных сегментов $\tilde{\alpha}_r$ определяются в соответствии с выражением (12).

Показатель $\hat{\alpha}$ описывает погрешность представления речевого сигнала в спектральной области с учетом особенностей восприятия речи человеком как некоторой совокупности формант. Он удовлетворяет перечисленным выше требованиям к показателю разборчивости, легко определяется по сегментам речевых сигналов с использованием хорошо разработанных алгоритмов линейного предсказания речи. В силу указанных свойств величина $\hat{\alpha}$ может быть использована для оптимизации систем обработки речевых сигналов с применением компьютеров.

2. Показатели натуральности речи

Под натуральностью переданной речи понимают ее свойство сохранять особенности произношения абонента. Натуральной может считаться такая речь, которая звучит естественно и позволяет узнавать диктора по голосу с высокой вероятностью.

В соответствии со сказанным выше в качестве одного из показателей натуральности речи может быть принята средняя вероятность ошибки автоматической верификации абонента $P_{ош.вер.}$ по переданному речевому сигналу заданной длительности. Отметим, что в теории опознавания речи [7] задача верификации дикторов (абонентов) состоит в принятии решения о том, принадлежит ли заданный фрагмент речи конкретному диктору, против альтернативы - фрагмент речи принадлежит другому человеку. Для ее оценки требуется наличие комплекса программ, реализующих алгоритмы автоматической верификации дикторов. С помощью такого комплекса программ среднюю вероятность ошибочной верификации дикторов можно оценить как выборочное среднее

$$\hat{P}_{ош.вер.} = k / n, \tag{15}$$

где k – число ошибочных решений о дикторе; n – общее число предъявленных на верификацию сегментов речевых сигналов.

Естественно, что чем больше значение показателя (15), тем выше натуральность речи.

Другой показатель натуральности принятой речи можно определить, исходя из точности передачи интонации произносимой речи. Интонация представляет собой систематическое изменение высоты звука на протяжении произносимого предложения. Она является важным аспектом речи, который содержит информацию о типе произносимых предложений, разбиениях и категориях фразовых структур, о моделях ударений, о семантике и эмоциональности. Высота вокализованных звуков характеризуется частотой основного тона, для оценивания которой разработано большое количество алгоритмов. Поэтому в качестве одного из показателей натуральности переданной речи может быть использовано среднеквадратическое отклонение частоты основного тона принятого сигнала от частоты основного тона переданного сигнала

$$\sigma_F = \sqrt{M[F - F_O]^2} \,. \tag{16}$$

Модифицируем показатель (16) к виду, удобному для оптимизации систем передачи речевых сообщений. Для этого представим сигнал на входе и на выходе системы передачи речи в виде совокупности следующих друг за другом коротких сегментов длительностью $T_C \approx 20$ мс. Тогда сигнал на входе системы передачи будет представлен совокупностью сегментов вида (11), а на выходе – (10).

Погрешность передачи высоты вокализованных звуков речи оценим величиной

$$\hat{\beta}_F = \frac{1}{\overline{F}_o} \sqrt{\frac{1}{P_v} \sum_{r=1}^{P_v} [\hat{F}_r - \hat{F}_{or}]^2} \,, \tag{17}$$

где $\overline{F}_o = \frac{1}{P_v} \sum_{r=1}^{P_v} \hat{F}_{or}$ является выборочным средним частоты основного тона вокализованных сегментов речевого сигнала; P_v – число вокализованных сегментов речи; $\hat{F}_{or}$ – оценка частоты основного тона для исходного сигнала, полученная на r-ом сегменте; $\hat{F}_r$ – оценка частоты основного тона для принятого сигнала, полученная на r-ом сегменте. Множитель $1/\overline{F}_O$ введен в (17) для получения значения показателя $\hat{\beta}_F$ в виде безразмерной величины. Оценивание частоты основного тона может быть выполнено по одному из алгоритмов, описанных в работах [3, 4, 6].

3. **Общая постановка задачи оптимизации систем передачи речевых сообщений**

Оптимизация систем передачи речи традиционно проводится путем сравнения различных вариантов систем с оценкой качества переданной речи на слух. При этом число опробованных вариантов невелико, и неизбежны ошибки за счет субъективности мнений экспертов. Для объективной оценки качества речи в ряде работ используется показатель формантой разборчивости, так как он может быть непосредственно рассчитан на основе формантной теории разборчивости речи [1,2]. Нам неизвестны работы, в которых формулировались и решались бы задачи автоматизированной оптимизации систем передачи речевых сигналов по сформулированным критериям качества.

Сформулируем постановку задачи оптимизации систем передачи речевых сообщений с применением введенных показателей качества речи.

Если оптимизация систем передачи проводится в процессе функционирования системы, то эта задача может быть отнесена к задачам адаптации.

В постановке задачи будем полагать, что структура системы передачи речевых сигналов известна, а неизвестными являются лишь некоторые параметры системы, образующие вектор $\vec{\gamma}$. Множество Γ_∂ допустимых значений этого параметра определяется конкретными особенностями прикладной задачи. Одним из важных ограничений, определяющих множество Γ_∂, является пропускная способность канала связи.

Качество передачи речи зависит от состояния системы и характеризуется показателем $k(\vec{\gamma})$, который в общем случае представляет взвешенную сумму частных показателей качества речи

$$k(\vec{\gamma}) = h_1\overline{\alpha} + h_2\hat{\beta}_F + h_3\hat{P}_{ош.вер.}, \qquad (18)$$

где значения показателя разборчивости $\overline{\alpha}$ определяется выражением (14), а значения показателей натуральности $\hat{\beta}_F$ и $\hat{P}_{ош.вер.}$ – выражениями (17) и (15) соответственно; весовые коэффициенты h_1, h_2, h_3 неотрицательны и характеризуют важность соответствующих частных показателей.

На вход системы передачи поступает исходный речевой сигнал $x_O(t)$, а с выхода снимается переданный сигнал $x(t)$.

Необходимо найти значение вектора параметров $\vec{\gamma}$ из условия минимума показателя $k(\vec{\gamma})$ по множеству допустимых значений Γ_∂

$$\vec{\gamma}_o = \arg\min_{\vec{\gamma}\in\Gamma_\partial}\{k(\vec{\gamma})\}. \qquad (19)$$

Отметим, что при соответствующем выборе весовых коэффициентов в выражении (18) качество речи может характеризоваться двумя или одним из частных показателей разборчивости и натуральности. Кроме того, вместо (18) могут применяться и другие методы сворачивания вектора показателей качества.

Использование описанного критерия качества позволяет сформулировать и решить ряд конкретных задач по оптимизации систем передачи речевых сообщений. Для решения таких задач можно использовать численные методы оптимизации, в частности градиентные методы и методы целочисленного программирования [9-10]. На заключительном этапе синтеза правильность полученного решения проверяется с привлечением бригады аудиторов. С применением предложенного подхода авторами статьи решен ряд задач по оптимизации низкоскоростных систем передачи речи. Полученные результаты планируется опубликовать в последующих выпусках журнала.

Заключение

В настоящей статье обоснованы показатели качества передачи речевых сообщений по цифровым каналам связи. Такими показателями является показатель разборчивости и показатели натуральности речи. Эти показатели приведены к виду, позволяющему автоматически вычислять их значения. Предложен подход к оптимизации алгоритмов передачи речи по цифровым каналам связи.

Список литературы: 1. *Покровский Н. В.* Расчет и измерение разборчивости речи. М.: Связьиздат, 1962. 390 с. 2. *Вемян Г.В.* Передача речи по сетям электросвязи. М.: Радио и связь, 1985. 272 с. 3. *Дж. Д. Маркел, А. Х. Грей.* Линейное предсказание речи. М.: Связь, 1980. 308с. 4. *Рабинер Л. Р., Шафер Р. В.* Цифровая обработка речевых сигналов /Под ред. М. В. Назарова и Ю. Н. Прохорова . М.: Радио и связь, 1981. 496 с. 5. *Марпл С.Л. (мл.)* Цифровой спектральный анализ и его приложения. М.: Мир, 1990. 584 с. 6. *Омельченко А.В., Пресняков А.И.* Статистический синтез алгоритмов оценивания периода основного тона речевых сигналов Радиоэлектроника и информатика. 1999. № 1, С 24-28. 7. *Рамишвили Г.С.* Автоматическое опознавание говорящего по голосу. М.: Радио и связь, 1981. 224 с. 8. *Петрович Н.Г., Каблукова М.В., Козленко Н.И.* Передача сигналов методом КИМ-ОФТ. М.: Связь, 1974. 112 с. 9. *Ковалев М. М.* Дискретная оптимизация (целочисленное программирование). М.: Изд-во БГУ, 1977. 190 с. 10. *Сергиенко И. В., Лебедева Т. Т., Рощин В. А.* Приближенные методы решения дискретных задач оптимизации. Киев.: Наук. думка, 1980. 274 с.

Харьковский государственный технический университет радиоэлектроники

Поступила в редколлегию 13. 03. 2001