В. М. Алюшин, С. В. Дворянкин

ВОССТАНОВЛЕНИЕ ГАРМОНИЧЕСКОЙ СТРУКТУРЫ ИСКАЖЕННЫХ РЕЧЕВЫХ ВОКАЛИЗМОВ ПОСРЕДСТВОМ ЦИФРОВОЙ ОБРАБОТКИ ИЗОБРАЖЕНИЙ ДИНАМИЧЕСКИХ СПЕКТРОГРАММ

Введение

Запись речи в естественных условиях часто осуществляется при высоком уровне шумов, что приводит к существенному снижению ее качества, в частности разборчивости. Использование для передачи речевых сообщений (РС) линий связи низкого качества или низкой пропускной способности приводит к возникновению дополнительных искажений и, как следствие, к существенному ухудшению их разборчивости, особенно на формантном и слоговом уровнях. Перечисленные обстоятельства обусловливают актуальность разработки эффективных методов шумоочистки РС [1–3], в первую очередь, на основе современных технологий цифровой обработки сигналов (ЦОС).

Применяемые на практике методы шумоочистки, как правило, базируются на анализе РС в спектральной области [4–6] и могут быть условно подразделены на две группы. Методы первой группы преимущественно используют простые алгоритмы анализа непродолжительных участков РС и характеризуются невысокой эффективностью [7]. Методы этой группы обычно применяют только для шумоочистки РС, дающей возможность повысить разборчивость без идентификации говорящего. Основным достоинством данных методов являются невысокие требования к необходимым вычислительным ресурсам.

Методы второй группы [8] позволяют осуществить полный анализ РС при высоком уровне шума и, в частности, решить задачу идентификации говорящего. Однако для их реализации требуются существенно большие вычислительные мощности.

Одним из наиболее перспективных подходов к очистке РС от шумов следует считать метод, основанный на восстановлении гармонической структуры речевых вокализмов, которые в дальнейшем используются для синтеза восстановленного речевого сообщения на основе обратного преобразования Фурье [9, 10].

Целью данной работы являлась разработка метода шумоочистки РС, предполагающего восстановление гармонической структуры искаженных речевых вокализмов посредством цифровой обработки изображений динамических спектрограмм.

Ключевую роль для восстановления гармонической структуры речевых вокализмов играет так называемая частота основного тона (ЧОТ), для определения которой в работе предлагается

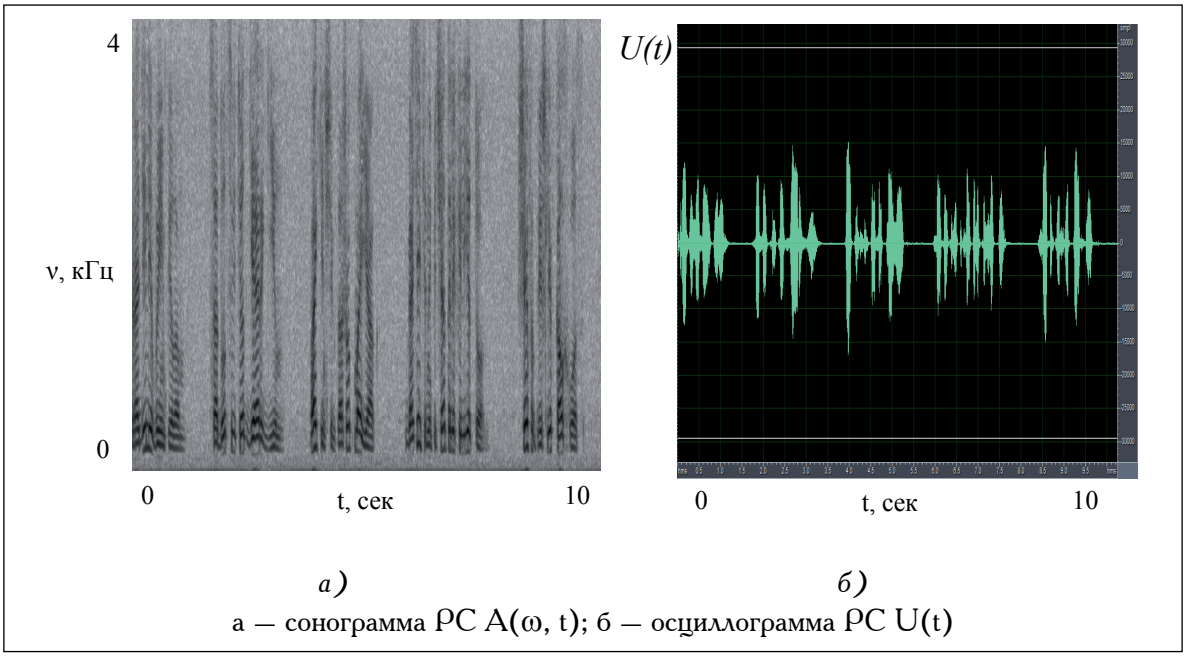
использовать спектральные характеристики РС. Основными требованиями, которые были предъявлены к разрабатываемому методу, являлись:

- высокая точность нахождения ЧОТ в условиях высокого уровня помех;
- возможность реализации метода в виде многопоточного приложения с целью наиболее полного использования вычислительных возможностей современных многоядерных процессоров, в частности видеопроцессоров;
- возможность обработки нескольких РС в режиме реального времени.

Спектральное описание РС

Как известно, смысловая часть информации содержится в частотной огибающей РС [11], а основой конструкций всех РС являются вокализованные участки речи [12, 13]. Изменением формы огибающей спектра речи достигается понятное и разборчивое звучание всего РС. В случае вокализованного участка речи, соответствующего произношению гласных звуков, спектр сигнала имеет гармоническую структуру, то есть представляет собой набор гармоник, все частоты которых кратны ЧОТ [9].

На рис. 1а изображена сонограмма РС $A(\omega, t)$ в оттенках серого цвета, а на рис. 1б представлена его осциллограмма $U(t)$. Интенсивность черного цвета каждой точки (пикселы) на сонограмме $A(\omega,t)$ пропорциональна логарифму амплитуды гармоники с частотой ω в момент времени t.

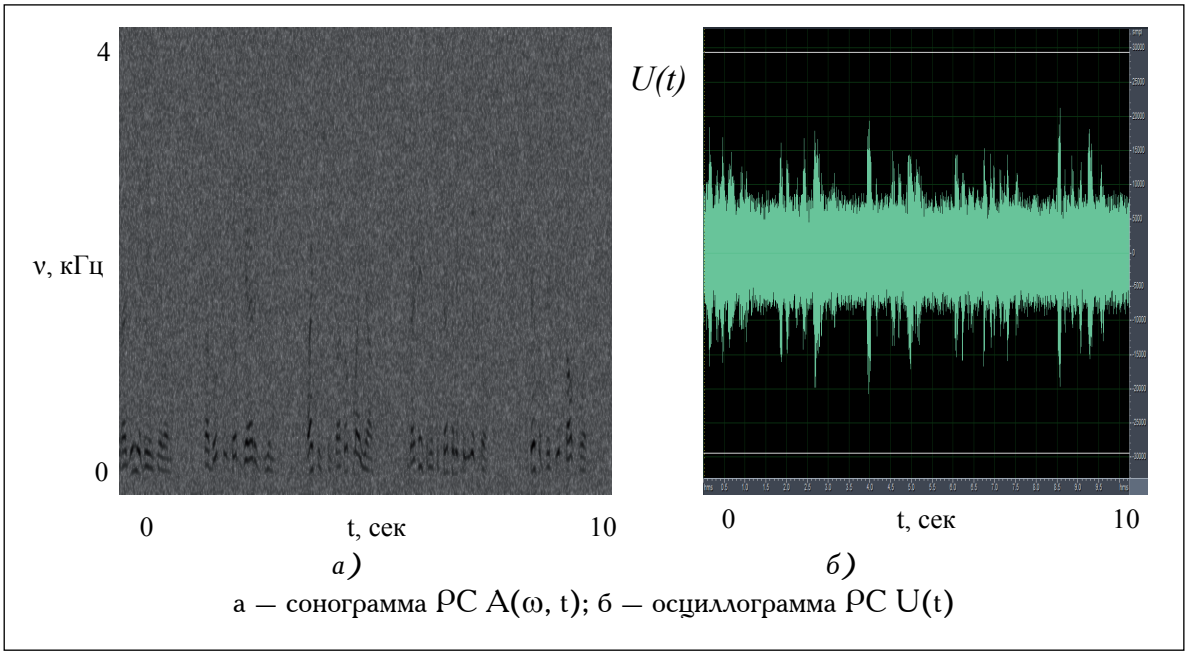


а — сонограмма РС A(ω, t); б — осциллограмма РС U(t)

Рис. 1. Спектральная и временная характеристики РС

Приведенные на рис. 1 спектральная и временная характеристики типичны для РС с невысоким уровнем шума. На рис. 2 представлены аналогичные характеристики для РС с высоким уровнем шума.

Как правило, в этом случае пикселы сонограммы РС, соответствующие высокочастотным (верхним) гармоникам, как наименее мощные и, следовательно, наименее яркие и контрастные, маскируются пикселами, соответствующими шумовым сигналам. Только пикселы, соответствующие нескольким мощным низкочастотным гармоникам в нижней части спектра с наибольшей амплитудой, проявляются на фоне «шумовых» пикселов.



а — сонограмма РС A(ω, t); б — осциллограмма РС U(t)

Рис. 2. Спектральная и временная характеристики РС с высоким уровнем шума

Было проведено исследование, направленное на выявление возможности точного определения ЧОТ на основе оставшейся значимой информации о гармониках, содержащейся в сонограмме, искаженной шумами, с целью восстановления всей гармонической структуры РС.

Метод нахождения ЧОТ по вершинам парабол гармоник узкополосных сонограмм

Указанный метод является следствием обобщения результатов, полученных с помощью различных подходов к определению ЧОТ на основе информации, взятой из изображений узкополосных сонограмм РС, искаженного шумами.

Один из исследуемых подходов предполагал использование в качестве начального приближения для ЧОТ $\omega_{\text{осн}}$ частоты гармоники сонограммы с самой большой амплитудой:

$$\omega_{\text{осн}} = \frac{2\pi}{N}\nu \cdot \arg \max_{i=0\ldots\frac{N}{2}-1} |X_i|,$$

где N — база преобразования Фурье; $\{X_i\}_{i=0}^{N-1}$ — кратковременное дискретное преобразование Фурье для вокализованного участка речи; ν — частота дискретизации исходного звукового сигнала.

Однако данному подходу свойственен ряд недостатков, ограничивающих область его применения:

1) На практике достаточно часто максимальной амплитудой обладает не основной тон, а одна из кратных ему гармоник нижней части спектра. Поэтому найденная таким образом ЧОТ может кратно превышать свое истинное значение.

2) ЧОТ может быть определена с погрешностью $\Delta\omega = \frac{2\pi}{N}\nu$ (разница между частотами двух соседних гармоник при дискретном преобразовании Фурье).

3) В каждый интервал времени кратковременного преобразования Фурье ЧОТ определяется независимо от других интервалов времени, что может приводить к увеличению ошибки.

4) Частоты всех гармоник на вокализованных участках речи пропорциональны ЧОТ, что приводит к росту погрешности их определения с увеличением номера гармоники. В итоге линии гармоник, определенных таким образом, получаются разрывными.

На рис. 3 представлен типичный результат восстановления набора гармоник (гармонической структуры искаженных речевых вокализмов) на основе определенного в рамках рассматриваемого подхода значения ЧОТ.

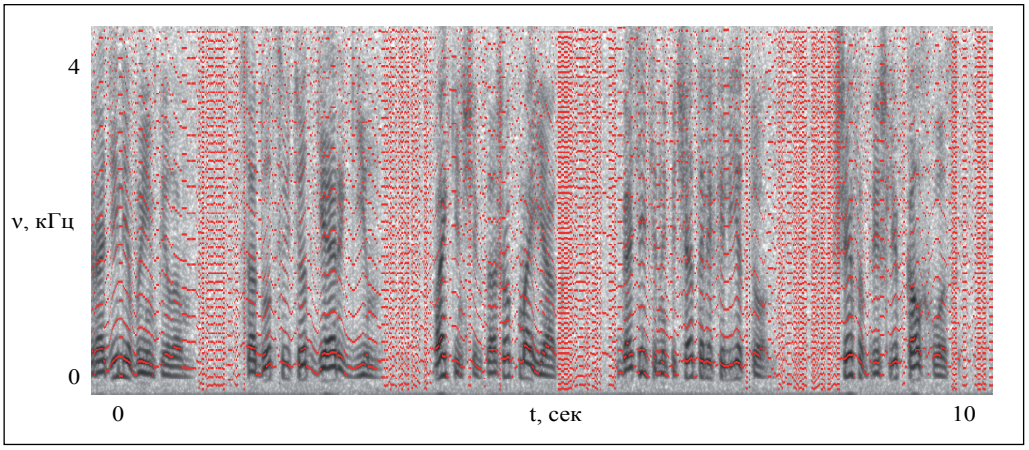


Рис. 3. Сонограмма и линии гармоник (красные), найденные без корректировки ЧОТ

В качестве второго подхода к определению ЧОТ был рассмотрен подход, предполагающий ограничение частотного диапазона ее поиска. В данном случае ЧОТ $\omega_{\text{осн}}$ определялась не на всем диапазоне частот, а только в области частот $\omega \in [70 - 750$ Гц], где с большой вероятностью находятся гармоники речевых вокализмов:

$$\omega_{\text{осн}} = \frac{2\pi}{N}\nu \cdot \arg \max_{i=\frac{70\cdot N}{\nu}\dots\frac{750\cdot N}{\nu}} |X_i|.$$

Основной проблемой реализации данного подхода на практике является возможность ошибочного определения ЧОТ по частоте второй или третьей гармоники, что может привести к ошибке в определении ЧОТ соответственно в два и три раза.

Для устранения указанной проблемы была разработана методика, позволяющая в значительной степени повысить точность определения ЧОТ, вычисляемой в соответствии со вторым подходом.

Пусть a – номер элемента X_a дискретного преобразования Фурье с максимальной амплитудой (в этом случае ЧОТ равна $\omega_{\text{осн}} = \frac{2\pi}{N}\nu \cdot a$). Рассмотрим элемент $X_{a/2}$. Если его амплитуда незначительно отличается от амплитуды элемента X_a $\left(\left||X_{a/2}| - |X_a|\right| < \frac{2}{3}|X_a|\right)$, то это свидетельствует о существовании значимой по энергии гармоники с меньшей частотой. Это возможно только тогда, когда за ЧОТ была ошибочно выбрана частота второй гармоники. Поэтому для определения ЧОТ необходимо уменьшить найденную частоту в два раза.

Аналогичным образом рассмотрим элемент $X_{a/3}$. Если его амплитуда незначительно отличается от амплитуды X_a,то этот факт однозначно свидетельствует о существовании гармоники с достаточно высоким уровнем энергии, частота которой в три раза ниже выбранной. Для определения ЧОТ в этом с лучае необходимо уменьшить выделенную частоту в три раза $\left(\left||X_{a/3}| - |X_a|\right| < \frac{2}{3}|X_a|\right)$. Данная методика позволяют исключить возможность ошибочного определения ЧОТ в нижней части спектра.

Третий исследовавшийся в работе подход был направлен на решение проблемы прерывистости высокочастотных гармоник, которые после их восстановления в соответствии с первым подходом плохо совпадают с исходными значениями. Это связано с тем, что ЧОТ определяется с точностью до величины $\Delta\omega = \frac{2\pi}{N}\nu$ (разница частот между двумя соседними по вертикали пикселами на

сонограмме) и данное расхождение между истинным и рассчитанным положением линии гармоники основного тона растет линейно с ее номером.

Для исправления этого недостатка был разработан метод определения ЧОТ по вершинам парабол гармоник узкополосных сонограмм. Сущность этого метода заключается в следующем.

Пусть X_a — элемент дискретного преобразования Фурье: локальный максимум, в первом приближении соответствующий основному тону. Построим параболу через три соседних элемента преобразования Фурье:

$$y = f(x) = Ax^2 + Bx + C$$

$$|X_{a-1}| = f(a-1),\ |X_a| = f(a),\ |X_{a+1}| = f(a+1).$$

Решая систему уравнений относительно A, B, C, получаем:

$$A = \frac{|X_{a+1}| - 2|X_a| + |X_{a-1}|}{2};$$

$$B = |X_a| - |X_{a-1}| - A \cdot (2a-1);$$

$$C = |X_a| - Aa^2 - Ba.$$

Найдем координату вершины параболы $x_0 = -\frac{B}{2A}$, тогда уточненное значение ЧОТ будет равно

$$\omega_{\text{осн}} = \frac{2\pi}{N}\nu \cdot x_0.$$

Следует заметить, что после осуществления операции коррекции найденное значение ЧОТ $\omega_{\text{осн}}$ оказывается не кратным шагу преобразования Фурье.

Сущность этого эффекта может быть объяснена с помощью следующего примера. Рассмотрим тестовый сигнал с частотой дискретизации $\nu = 8$ кГц, состоящий из двух гармоник:

$$y(t) = A_1 \cos(\omega_1 t + \varphi_1) + A_2 \cos(\omega_2 t + \varphi_2),$$

где $\omega_1 = 2\pi\nu$, $\omega_2 = 4\pi\nu$, $\nu = 503$ Гц (частота основного тона).

Для построения каждого спектрального среза будем использовать быстрое преобразование Фурье (БПФ) с базой N. Осуществим фильтрацию 1024 входных отсчетов РС, используя для этой цели рекомендуемое в [12] окно Гаусса:

$$w(x) = e^{-\frac{(x-N)^2}{2N^2\sigma^2}},$$

где x — номер отсчета ($0 \le x < N$), $\sigma = 0{,}17$.

Применив преобразование Фурье к данной последовательности, получаем развертку спектральных срезов, показанную на рис. 4.

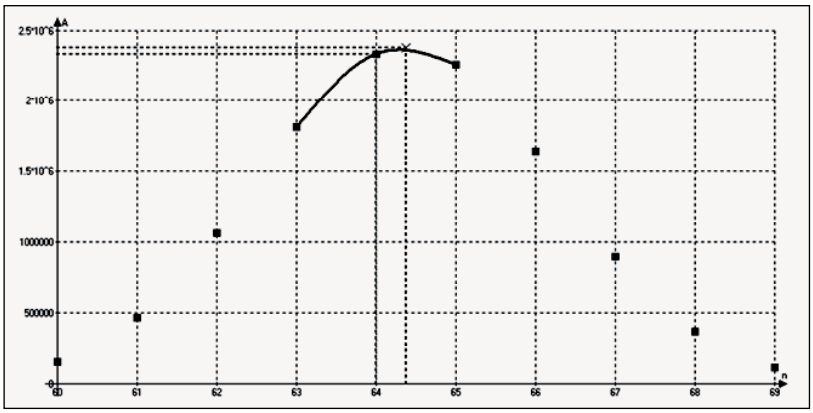


Рис. 4. Развертка спектральных срезов тестового сигнала, состоящего из двух кратных гармоник 503 Гц и 1006 Гц

Наибольшей амплитудой (2331229,000) обладает гармоника № 64. Частота данной гармоники $\nu_{64} = \frac{\nu}{N} n = \frac{8000}{1024} \cdot 64 = 500$ Гц, что отличается от истинной частоты основного тона на 3 Гц. Несмотря на то что полученная ошибка достаточно мала, она будет приводить к линейному росту погрешности при определении частоты гармоник в верхней части спектра. Так, например, погрешность определения частоты 10-й гармоники составит уже 30 Гц.

Произведем корректировку значения ЧОТ с помощью предлагаемого метода. Рассмотрим две соседние гармоники: № 63 с амплитудой 1812567,875 и № 65 с амплитудой 2255501,750. Проведем через них параболу и вычислим по описанным выше формулам коэффициенты: $A = -297194{,}1875; B = 38262322{,}9375$. Координата вершины параболы $x_в = -\frac{B}{2A} = 64{,}3726$, что соответствует частоте $\nu_в = \frac{\nu}{N} x_в = \frac{8000}{1024} \cdot 64{,}3726 = 502{,}9109$ Гц.

Таким образом, применение описываемого метода в рассматриваемом случае дает возможность восстановить значение ЧОТ с ошибкой менее чем в 0,1 Гц.

На рис. 5 представлены результаты определения частоты гармоник с помощью рассматриваемого метода. Как видно, после проведенной коррекции треки даже верхних гармоник являются непрерывными и совпадают с точными исходными значениями на изображении узкополосной спектрограммы.

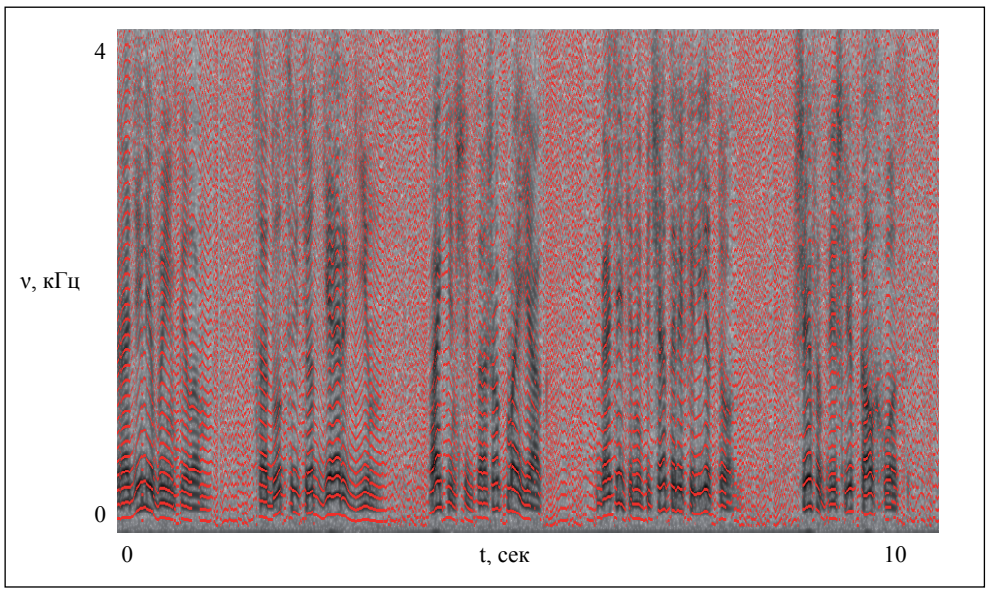


Рис. 5. Сонограмма и линии гармоник (красные) после коррекции ЧОТ с помощью предложенного метода

В качестве дополнительного критерия правильности определения ЧОТ предлагается использовать такую характеристику, как максимальная сумма амплитуд первых 7 кратных гармоник:

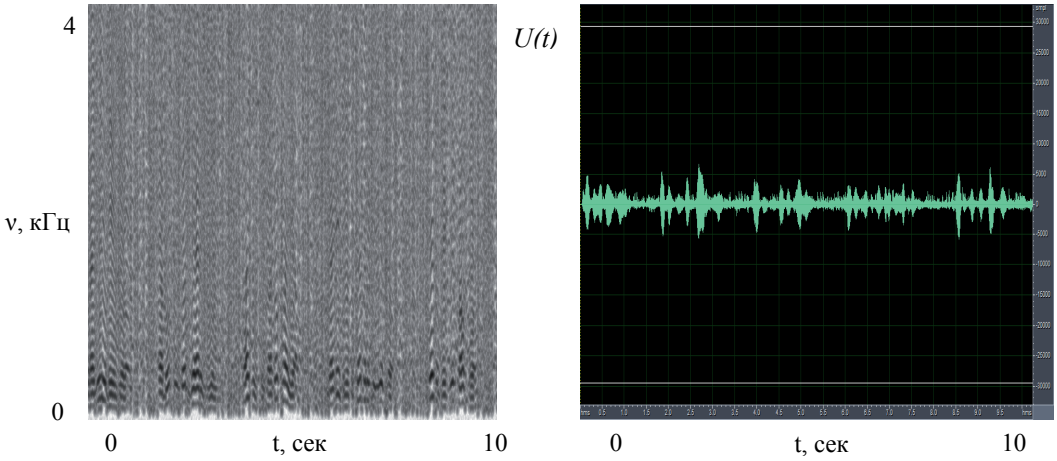
$$x_в = \arg \max_{x \in \left[\frac{70 \cdot N}{\nu}; \frac{750 \cdot N}{\nu}\right]} \sum_{i=1}^{7} \left| X_{[i \cdot x]} \right|,$$

где $[i \cdot x]$ — целая часть числа $i \cdot x$.

Результаты учета этого критерия продемонстрированы на рис. 5.

Как показало экспериментальное исследование метода, он применим даже для восстановления РС с высоким уровнем шума, при котором на сонограмме видны только некоторые первые гармоники на фоне шумов (рис. 2а). На рис. 6 изображена сонограмма и осциллограмма РС

после осуществления процедуры шумоочистки с помощью предложенного метода. Для анализа был рассмотрен зашумленный РС продолжительностью 10 с при частоте дискретизации 8 кГц.



а) *б)*

а — сонограмма РС A(ω, t); б — осциллограмма РС U(t)

Рис. 6. Спектральная и временная характеристики РС после восстановления гармонической структуры вокализмов

Для оценки качества работы метода было рассмотрено отношение сигнал/шум, рассчитанное при помощи пик-фактора. Так, для исходного РС оно равнялось 0,60, а после шумоочистки — 0,84.

Как показали эксперименты, при наличии в РС белого шума треки гармоник и гармоническая структура вокализмов восстанавливаются корректно даже при отношении сигнал/шум Δ до −12 дБ. При наличии в РС естественных помех гармоническая структура вокализмов может быть восстановлена корректно при отношении сигнал/шум до −8…−5 дБ, что обуславливает эффективность разработанного метода.

Для реализации предложенного метода определения ЧОТ вокализованного участка зашумленной речи была разработана программа на языке C++ с использованием технологии NVIDIA CUDA. Программа ориентирована на использование ПК либо ноутбука с любой видеокартой NVIDIA, графический процессор которой содержит не менее 236 вычислительных ядер. Программа позволяет осуществить в режиме реального времени обработку до 8 РС с произвольной частотой дискретизации, предполагающую восстановление их гармонической структуры и шумоочистку в соответствии с предложенным методом.

Возможности учета персональных характеристик диктора

После нахождения основного тона искаженного речевого сообщения возможно восстановление его гармонической структуры на каждом временном срезе спектрограммы с использованием следующего соотношения:

$$\omega_i = i \cdot \omega_{осн}, i \in \left[1; \frac{\nu}{2 \cdot \omega_{осн}}\right],$$

где ω_i — круговая частота i-й гармоники.

По полученной гармонической структуре возможно построение скорректированной спектральной огибающей улучшенного РС конкретного диктора с использованием накопленной базы данных его речевых вокализмов в результате совмещения новой гармонической и взятой из заранее созданной голосовой базы связанной с ней формантной структуры. Найденная таким образом

скорректированная спектральная огибающая может использоваться для решения задач, связанных с синтезом акустического сигнала с заданными свойствами, восстановлением разборчивости РС, идентификацией личности говорящего человека и др. В частности, для идентификации диктора и распознавания элементов речи могут использоваться такие характеристики РС, как частота основного тона, тембр голоса, характерные черты формантной структуры речи и др.

Применение разработанного метода нахождения ЧОТ для восстановления формантной структуры речи

Найденную гармоническую структуру целесообразно применить для восстановления формантной структуры речи с использованием базы данных голоса диктора, в которой хранятся записи отдельных фраз, а также изображения соответствующих им сонограмм. Для этого, прежде всего, необходимо разделить РС на отдельные вокализмы, слога или слова. В качестве границы слов можно использовать паузы, в которых отсутствует гармоническая структура или амплитуда гармоник мала. После этого необходимо для каждого вокализма, слога или слова построить изображение частотной огибающей по тем гармоникам, которые наиболее заметны на фоне шумов. Затем необходимо в базе данных вокализмов, слогов или слов диктора найти изображение, наиболее похожее на каждое полученное изображение частотной огибающей. Далее в очищенном РС следует заменить каждую фразу на запись из базы данных голоса. В этом случае из РС полностью удалятся все помехи, но следует как можно точнее находить подходящий вокализм, слог или слово в базе данных. Для этого можно сравнивать изображения частотных огибающих с использованием метрики Минковского:

$$\varepsilon_2 = \frac{1}{NM}\left\{\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}\left|C(i,j)-\hat{C}(i,j)\right|^2\right\}^{1/2},$$

где N и M — высота и ширина изображений соответственно, $C(i,j)$ и $\hat{C}(i,j)$ — яркости (от 0 до 255) пиксела с координатами (i, j) на двух сравниваемых изображениях.

Разделение РС на отдельные слова, а также процедуру сравнения изображений можно осуществить с помощью специализированного ПО Sound Tool.

Альтернативным способом восстановления формантной структуры РС является применение специализированного ПО для фрагментации РС на отдельные слова, а затем использование стандартных средств распознавания изображений (OCR), например, ABBYY Fine Reader для поиска в БД наиболее подходящей фразы.

Заключение

В настоящее время актуальна задача шумоочистки искаженного РС и получения из него исходной смысловой информации. Одним из способов нейтрализации помех является восстановление гармонической структуры речи. В данной статье предложен метод определения ЧОТ вокализованного участка зашумленной речи по вершинам парабол гармоник на спектральных развертках. Разработанное ПО, реализующее данный метод, позволяет в режиме реального времени при небольших вычислительных мощностях восстанавливать гармоническую структуру сигнала, на основании которой с помощью голосовой базы данных диктора можно восстановить связанную с ней формантную структуру, что является основой для решения таких актуальных в настоящее время задач, как восстановление разборчивости РС, идентификация диктора, синтез речи по изображению сонограмм, очистка речи от шумов и помех, сжатие и восстановление РС.

СПИСОК ЛИТЕРАТУРЫ:

1. *Максимов Е. М., Ромашкин Ю. Н., Лопатина С. А.* Актуальные задачи речевой акустики // Речевые технологии. 2008. № 2. С. 66–70.
2. *Хитров М. В.* Распознавание русской речи: состояние и перспективы // Речевые технологии. 2008. № 1. С. 83–87.
3. *Михайлов В. Г.* Из истории исследований преобразования речи // Речевые технологии. 2008. № 1. С. 93–113.
4. *Колоколов А. С., Павлова М. И.* Способ обработки речевого сигнала в частотной области. Патент РФ № 2454735. 2006.
5. *Пэй У.* Частотный фильтр и способ фильтрации в частотной области. Патент РФ № 2308153. 2006.
6. *Ронгшан Юй Браун, Филлип С.* Повышение разборчивости речи с помощью четкости голоса. Патент РФ № 2469423, 2008.
7. *Азаров И. С., Петровский А. А.* Вычисление мгновенных гармонических параметров речевого сигнала // Речевые технологии. 2008. № 1. С. 67–77.
8. *Жиляков Е. Г., Курлов А. В., Эсауленко А. В., Котович Н. В.* Об одном методе очистки речи от шумов на основе применения фильтрующей субполосной матрицы // Доклады 11-й Международной конференции DSPA-2011. Обработка сигналов в системах телекоммуникаций. С. 197– 200. http://www.autex.spb.ru/dspa2011.php
http://www.autex.spb.ru/dspa/dspa2011-3-2.doc
9. *Рабинер Л. Р., Шафер Р. В.* Цифровая обработка речевых сигналов: Пер. с англ. / Под ред. М. В. Назарова и Ю. Н. Прохорова. М.: Радио и связь, 1981. – 496 с.
10. *Дворянкин С. В.* Цифровая шумоочистка аудиоинформации / Под ред. д.т.н., профессора А. В. Петракова. М.: ИП РадиоСофт, 2011. – 208 с.
11. *Калинцев Ю. К.* Разборчивость речи в цифровых вокодерах. М.: Радио и связь, 1991. – 218 с.
12. *Кузнецов В. Б., Чучупал В. Я.* Классификация звуков русской речи с помощью бинарных решающих деревьев // Речевые технологии. 2008. № 2. С. 24–35.
13. *Женило В. Р.* Компьютерная фоноскопия. М.: Изд-во Акад. МВД России, 1995. – 208 с.