УДК 534.441

ГИБРИДНЫЙ ДЕТЕКТОР РЕЧИ

Вознесенская Т.В., *к.ф-м.н. доцент Департамента больших данных и информационного поиска Факультета компьютерных наук НИУ ВШЭ, e-mail: tvoznesenskaya@hse.ru;*
Котов М.А., *руководитель научно-технического департамента ООО «Стел-Компьютерные Системы», e-mail: kotov@stel.ru;*
Леднов Д.А., *к.т.н., с.н.с, научный консультант научно-технического департамента ООО «Стел-Компьютерные Системы», e-mail: lednov@stel.ru.*

Ключевые слова: детектор речи, гибридный, звуковой поток, теория обнаружения, фрейм, линейчатый спектр, тестирование.

Рассматривается гибридный детектор речи, который построен из двух последовательно соединенных детекторов, обладающих различными принципами работы. Показано, что он способен выделять речь на фоне нестационарных шумов, обладающих сплошным спектром, при низких соотношениях сигнал/шум. Однако, такой детектор не способен отличить речь от речеподобных сигналов, обладающих линейчатыми спектрами, например, от музыки. Для решения этой задачи был использован известный детектор, основанный на вычислении отношения правдоподобия статистических моделей музыки и речи, полученных в процессе обучения. Приводятся экспериментальные результаты исследования работы гибридного детектора.

Введение

Детектор речи предназначен для выделения из входного звукового потока, состоящего из смеси полезного (речевого) сигнала и шума, последовательности сегментов, каждый из которых содержит фразу или слово. Задача обнаружения речи сходна с задачей, решаемой в рамках классической теории обнаружения стохастических сигналов и описанной во многих классических публикациях (см. например [1, 2]). Основное положение этой теории состоит в том, что риск принять неправильное решение минимален, если решение принимается в виде

$$\delta = \begin{cases} 1, & \text{если } \lambda > \mu \\ 0, & \text{если } \lambda \le \mu \end{cases}, \quad (1)$$

где решение $\delta = 1$, соответствует решению, что сигнал содержит речь, а решение $\delta = 0$, соответствует решению, что сигнал содержит только шум. В формуле (1) введены обозначения:

$$\mu = \frac{(1-p_1)K(0,\delta_1)}{p_1 K(1,\delta_0)},$$

$$\lambda = \frac{p(y\,|\,1)}{p(y\,|\,0)}.$$

Отношение λ как правило, называют отношением правдоподобия, y – наблюдаемый сигнал, $p(y\,|\,\theta)$ – условная плотность распределения вероятности (ПРВ) наблюдаемого сигнала в зависимости от случайной величины $\theta = \{0,1\}$, p_1 – априорная вероятность наличия полезного сигнала в наблюдаемом процессе, $K(\theta,\delta)$ – положительно определенная функция потерь, выбираемая эмпирически, как правило, в виде

$$K(\theta,\delta) = K_{\theta\delta} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Реализация данной схемы для детектирования речи приводит к тому, что необходимо иметь представление о виде ПРВ $p(y\,|\,\theta)$. Поскольку знанием об истинной форме ПРВ речи и шума мы не обладаем, то для решения этой проблемы обычно предполагается, что данную ПРВ можно аппроксимировать гауссовой смесью [3] вида

$$p(y\,|\,\theta) = \sum_{i=1}^{n} \alpha_i p_i(y\,|\,\Phi_i,\theta),$$

где $p_i(y\,|\,\Phi_i,\theta)$ – нормальная ПРВ с параметрами Φ_i, в качестве параметров используются ковариационная матрица и вектор математических ожиданий, α_i – априорная вероятность нормальной ПРВ, n – количество элементов смеси.

Следующий шаг, это вычисление параметров $\{\Phi_i\,|\,\theta\}$ гауссовой смеси. Для этого необходимо создать два множества записей – одно множество будет содержать образцы речи различных дикторов, а другое множество образцы различных типов шумов, а затем на основе собранного материала, используя известный EM-алгоритм [3], найти неизвестные параметры.

Сложность описанного алгоритма состоит в том, что нет практической возможности собрать образцы всех возможных типов неречевых сигналов, чтобы построить их статистические модели (плотности распределения вероятностей). В свою очередь это приводит к росту числа ошибок в работе детектора речи, как только требуется выделить речь на фоне подобных неречевых сигналов.

Возникает вопрос: может ли быть построен детектор, работа которого основана на некотором устойчивом признаке речи, отличающим ее от всех прочих шумов?

В настоящей работе в качестве такого устойчивого признака используем то, что речь обладает вокализованными интервалами, т.е. при произнесении каждого слога произносится гласный звук, спектр которого обладает признаками линейчатости [4]. Предположительно, этот подход позволит нам выделить речь на фоне широ-

кого класса шумов обладающих сплошным спектром, при этом шумы могут иметь даже нестационарный характер.

Бесспорно, признаком линейчатости спектров не обладает «шепотная» речь, и в противовес, этим признаком обладают музыкальные произведения. Что касается «шепотной» речи, то она не представляет практического значения и не встречается в сообщениях, требующих обработки. Что касается музыкальных произведений, то они будут являться источником ошибок ложных вызовов такого детектора. Однако, опыт практического использования детектора, основанного на статистических признаках, показывает, что он успешно справляется с задачей детектирования музыки. Это наводит на мысль о создании гибридного детектора речи, включающего в себя использование, как статистических, так и детерминированных характеристик речи.

В следующих разделах будут последовательно описаны методы выделения детерминированных характеристик и принципы построения детектора речи, основанного на этих характеристиках, а так же принципы построения гибридной модели детектора. В заключении будут приведены экспериментальные данные работы гибридного детектора.

Метод выделения линейчатой части спектра вокализованного звука

В качестве модели вокализованного звука используем следующее представление [4]:

$$x(t) = \int d\rho \sum_{i=1}^{m} K(\omega_i, \rho)\cos(\rho t + \phi(\rho)), \quad (1)$$

где в качестве ядра модели выбрано выражение

$$K(\omega_i, \tau_i, \rho) = A_i \exp\left(-\frac{(\omega_i - \rho)^2}{\tau_i^2}\right), \quad (2)$$

где τ_i ω_i – ширина и частота i-ой гармоники частоты основного тона, соответственно; $\phi(\rho)$ – фаза частоты ρ.

Такая модель сигнала позволяет нам ввести понятие ширины i-го обертона линейчатого спектра τ_i, которая всегда наблюдается при измерении спектра реального речевого сигнала. Причины возникновения ширины обертона могут быть различные: с одной стороны, это может быть изменение частоты обертона за время проведения измерений, а с другой стороны, длительность измерений может быть не кратна периоду обертона.

Для анализа спектра, заданного представлением (1), (2), воспользуемся вейвлет-преобразованием в частотной области [5]:

$$L(\omega, \tau, \{\omega_i, \tau_i\}) = \sum_{i=1}^{m} \theta(\omega, \tau, \omega_i, \tau_i) = \\ = \sum_{i=1}^{m} \int_0^{\Omega} \psi(\rho, \omega, \tau) K(\rho, \omega_i, \tau_i) d\rho, \quad (3)$$

где [0, Ω] – диапазон спектра, $S(\rho, \{\omega_i\})$ – спектр Фурье вокализованного звука со своим множеством гармоник, $\psi(\rho, \omega, \tau)$ – вейвлет-функция, в качестве которой выбрана модифицированная «мексиканская шляпа»

$$\psi(\rho, \omega, \tau) = \frac{1.031}{\sqrt{2}\tau^{3/2}} \exp\left(-\frac{(\omega - \rho)^2}{\tau^2}\right)\left(1 - 2\frac{(\omega - \rho)^2}{\tau^2}\right), \quad (4)$$

где τ – масштаб вейвлет-функции.

Численное интегрирование показывает, что если спектр состоит из одного обертона ω_1, ширина которого τ_1, то максимум интеграла

$$(\omega_1, \tau_1) = ind \max_{\omega, \tau} \theta(\omega, \tau, \quad \omega_1, \tau_1)$$

соответствует положению и ширине обертона.

Этот вывод позволяет нам выполнить следующие операции:

– в диапазоне частот спектра $[\Omega_1, \Omega_2]$ для каждого значения частоты и каждого значения ширины обертона вычислить значения интегралов $\theta(\omega, \tau)$ (диапазон значений ширин обертонов выбирается из следующих соображений: ширина обертона не может быть менее, чем четверть минимальной частоты основного тона, и более, чем четверть максимальной частоты основного тона);

– для каждого значения частоты основного тона ω_0 рассчитать значение суммы

$$F(\omega_0) = \sum_{i=1}^{m} I(i)\theta(i\omega_0), \quad (5)$$

где индикаторная функция

$$I(i) = \\ = \begin{cases} 1, \; if \;\; \theta(i\omega_0) > 0 \;\; and \;\; (\;\theta((i-1)\omega_0) > 0 \;\; or \;\; \theta((i+1)\omega_0) > 0), \\ 0, \qquad else \end{cases}$$

применение которой позволяет избежать эффектов принятия за частоту основного тона единичных спектральных всплесков, и найти величину

$$\theta(i\omega_0) = \max_{\tau} \theta(i\omega_0, \tau);$$

– разделить диапазон частот основного тона (90 Гц-450 Гц) на три непересекающиеся области (90 Гц-179 Гц, 180 Гц-359 Гц и 360 Гц-450 Гц) и для каждой области найти максимальное значение суммы (5) и частоту основного тона, которая доставила сумме этот максимум, т.е.

$$F(\omega^*) = \max_{\omega_0} F(\omega_0),$$

$$\omega^* = \arg\max_{\omega_0} F(\omega_0).$$

Таким образом, результатом выполненных операций будет множество параметров $\left\{\omega_i^*, F(\omega_i^*), \boldsymbol{\theta}(\omega_i^*)\right\}_{i=1,..,3}$, вычисленное для каждого диапазона частоты основного тона в диапазоне спектра $[\Omega_1, \Omega_2]$.

Сделаем пояснения к изложенной последовательности операций. Известно, что методы выделения мгновенной частоты основного тона страдают ошибками ее удвоения. Этот эффект возникает при условии, что амплитуды четных обертонов многократно превосходят амплитуды нечетных обертонов основного тона [6]. Эффект можно ослабить, если разделить диапазон частот основного тона на три непересекающиеся области (границы которых были указаны выше) и проводить обработку речи

в каждой из выделенных областей независимо. Поскольку эффект удвоения частоты является неустойчивым во времени, совместная обработка некоторой последовательности фреймов, каждый из которых обработан с учетом множества диапазонов частот основного тона, поможет снизить этот эффект. Области выбраны так, чтобы каждая из них не содержала удвоенных частот.

Пусть найдены множества параметров для последовательности из M фреймов. Введем меру того, что частота основного тона ω_{it}^{*}, полученная в i-ом диапазоне t-ого фрейма, имеет свое продолжение в j-ом диапазоне (t+1)-ого фрейма

$$\mu_{ijt} = \ln(F(\omega_{jt+1}^{*})) - \frac{(\omega_{it}^{*} - \omega_{jt+1}^{*})^2}{\sigma^2}, \quad (6)$$

где σ – допустимое изменение частоты основного тона от фрейма к фрейму (параметр модели). Тогда мера того, что за M фреймов частота основного тона двигалась по траектории по диапазонам с известными номерами $\{i_1,..,i_t,..,i_M\}$,будет равна

$$\beta(i_1,..,i_t,..,i_M) = \sum_{i=1}^{M-1} \mu_{i_t i_{t+1} t}. \quad (7)$$

Задача состоит в том, чтобы среди всех возможных траекторий найти траекторию с максимальной мерой (7). Для решения этой задачи можно использовать метод динамического программирования, подробно описанный в [7]. Заметим, что последовательность номеров диапазонов основного тона, которые определены методом динамического программирования для максимума меры (7) при условии, что значение этого максимума превышает некоторый порог Q (параметр модели), т.е.

$$\max_{i_1,..,i_t,..,i_M} \beta(i_1,..,i_t,..,i_M) > Q, \quad (8)$$

определяет саму частоту основного тона и множество амплитуд ее обертонов $\{\omega,\boldsymbol{\theta}\}$. Если же максимум меньше порога, то принимается решение, что основного тона в данной последовательности фреймов нет.

Итак, выполнение неравенства (8) приводит к тому, что принимается решение о наличии в каждом фрейме из последовательности из M фреймов вокализованного звука. Далее происходит смещение интервала анализа последовательности, состоящей из M фреймов на один фрейм, и вновь полученная последовательность проходит анализ в соответствии с формулами (6-8). Здесь интересны два случая: 1) до некоторого последнего фрейма в последовательности с номером n_t неравенство (8) не выполнялось, а в следующий момент времени n_{t+1} неравенство (8) уже выполняется; 2) (случай обратный первому) до некоторого последнего фрейма в последовательности с номером n_t неравенство (8) выполняется, а в следующий момент времени n_{t+1} неравенство (8) не выполняется. Первый случай соответствует событию начала вокализованного звука – в качестве этого момента времени нами будет приниматься момент времени соответствующего началу фрейма с номером n_{t+1-M}. Второй случай соответствует событию окончания вокализованного звука – в качестве этого момента времени нами будет приниматься момент времени соответствующий окончанию фрейма с номером n_t.

По сути, описанный метод дает нам значение длительности вокализованных звуков.

Структура гибридной модели

Известно, что каждый слог речи обязательно включает в себя вокализованный звук. В работе [8] подробно описаны характерные для речи частоты огибающих слогов. Если использовать низкочастотный фильтр с частотой среза равной средней частоте огибающей слогов при нормальном темпе речи (4 Гц), то можно сегментировать исходный звук на интервалы, предполагая, что каждый такой интервал включает в себя слог.

В качестве первичной обработки слога используем детектор, основанный на детерминированных характеристиках, который позволяет получить длительность вокализованной части слога внутри полученного интервала. Вычисляя отношение длительности вокализованной части d_v к длительности интервала D_i, мы определяем правдоподобность предположения, что внутри полученного интервала лежит слог. Если выполняется неравенство $\frac{d_v}{D_i} > Q_i$, где Q_i – параметр модели, то принимается решение о наличии слога на полученном интервале.

Вторичная обработка интервала производится только в том случае, если на первом этапе было принято положительное решение о звучании слога. Вторичная обработка состоит в вычислении отношения правдоподобия вида

$$\lambda = \prod_{t=1}^{R} \frac{p(x_t \mid speech)}{p(x_t \mid music)},$$

где R – количество фреймов в полученном интервале, $p(x_t \mid speech)$, $p(x_t \mid music)$ – условные вероятности того, что полученные в момент времени t характеристики фрейма x_t порождены речью или музыкой соответственно. В качестве характеристик фрейма нами были выбраны известные коэффициенты (melfrequencycepstralcoefficients) MFCC [9, 10].

Если выполняется неравенство $\lambda > Q_\lambda$, то принимается решение, что в интервале звучит речь, в противном случае – музыка.

Экспериментальные результаты

Прежде чем описывать экспериментальные результаты приведем технические характеристики системы, условия ее подготовки к экспериментам, а так же опишем тестирующую базу звуков.

Входной сигнал был оцифрован с частотой 8 кГц и имел разрядность – 16 бит. Для вычисления быстрого преобразования Фурье, которое как часть входит и в процедуру обнаружения линейчатых спектров и в вычисление коэффициентов MFCC, использовалось окно длительностью 64 мс, величина смещения окна во времени составляла 10 мс. В качестве фильтра низкой частоты использовался фильтр Баттерворта 5-го порядка. Частота среза фильтра ω_F была параметром модели. Для по-

Таблица 1

Соотношение сигнал/шум (дБ)	Значение равновероятной ошибки необучаемого детектора(%)	Значение равновероятной ошибки обучаемого, статистического детектора при распознавании речи и музыки (%)
10-15	9.86	1.08
6-10	19.435	4.43
2-6	29.185	9.12

строения статистических моделей речи и музыки использовались гауссовы смеси, состоящие из 16-ти элементов.

Для поиска параметров детерминированной модели, основанной на выделении вокализованных звуков, а так же для ее тестирования была создана речевая база, содержащая речь на фоне шумов КВ-канала радиоприемника, промышленных шумов (работа отбойного молотка, экскаватора, шум проезжающих машин, шум в вагоне метро и т.д.), бытовых и естественных шумов (шаги, работа вентилятора, шум листвы деревьев и т.д.). База была размечена, т.е. были установлены границы интервалов, включающих в себя речь.

Речевая база была разделена на две части, одна из которых использовалась для оптимизации параметров модели, а другая для тестирования. Длительность базы, которая использовалась для оптимизации параметров, составляла 1 час, она включала в себя 36 минут речи. Критерием оптимизации было минимальное значение суммы ошибок пропуска цели и ложного срабатывания. Приведем значения параметров, полученных в результате оптимизации: σ(допустимое изменение частоты основного тона от фрейма к фрейму) = 40 Гц; Q (порог меры (7))= 0,31; Q_i (отношение длительности вокализованной части слога к длительности слога) = 0.21.

Для тестирования использовалась база длительностью около 6 часов, которая включала в себя около 2-х часов речи. Выделенная детерминированным детектором из этой базы речь использовалась для создания ее модели, которая используется статистическим детектором. Для обучения статистической модели музыки был создан музыкальный корпус, включающий в себя около 8-и часов фрагментов музыкальных произведений различных темпов и стилей. Для тестирования статистической части детектора использовались размеченные записи, содержащие речевые и музыкальные фрагменты, каждый из таких фрагментов идентифицировался. Общая длительность записей составляла около полутора часов, из них около 20 минут музыки и около 30 минут речи.

Во втором столбце табл. 1 показаны значения равновероятной ошибки обнаружения речи в зависимости от соотношения сигнал/шум в поданных на вход необучаемого детектора записях. В третьем столбце табл. 1 показаны значения равновероятной ошибки классификации речевых и музыкальных фрагментов в зависимости от соотношения сигнал/шум в поданных на вход статистического детектора записях.

Заключение

По сути, созданный гибридный детектор, это смесь детерминированного детектора и системы распознавания. В настоящей работе были использованы только два класса распознавания: речь и музыка, но в общем случае количество классов может быть увеличено. В качестве классов могут быть, например, использованы голоса дикторов, которые нас не интересуют, или же языки, которые нас не интересуют и т.д.

Точность, полученная при высоких соотношениях сигнал/шум, сравнима с точностью, с которой одну и ту же запись сегментируют два эксперта. Поскольку каждый файл, использованной для тестирования базы, был обработан одним экспертом, то мы не могли собрать достоверную статистику в попытке ответить на вопрос, насколько наш детектор отличается от средней экспертной разметки. Анализ разницы между экспертной и детектированной разметками в основном состоит из разницы в определении границ слов и фраз, а также в расстановке пауз между словами. Эксперт имел склонность выделять слитную фразу целиком в отличие от автомата, который аккуратно расставлял паузы между словами.

При разработке детерминированного детектора было сделано допущение, которое существенным образом повлияло на его работу. В формуле (6) явным образом предполагается, что частота основного тона не должна изменяться от фрейма к фрейму, именно такая ситуация приведет к максимуму функционал (7), хотя практика показывает, что частота основного тона не является устойчивой. Это в свою очередь говорит о том, что должна быть разработана более совершенная модель, которая учитывает естественную неустойчивость частоты основного тона.

Литература

1. Большаков И.А. Статистические проблемы выделения потока сигналов из шума // Изд-во «Советское радио», 1969.

2. Харкевич А.А. Борьба с помехами, Изд. Второе, Из. «Наука», М. 1965.

3. Jeff A. BilmesA Gentle Tutorial of the EM algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models// ICSI Technical Report-021, April 1998.

4. Аграновский А.В., Леднов Д.А. Репалов С.А. Метод текстонезависимой идентификации диктора на основе индивидуальности произношения гласных звуков // Акустика и прикладная лингвистика. Ежегодник РАО. Выпуск 3. М., 2002, стр. 103-115

5. Котов М.А., Леднов Д.А. и др. Способ определения параметров линейчатых спектров вокализованных звуков и система для его реализации № 2007148606/09(053252) от 27.12.2007

6. Babkin A.A. LPC Speech Coder AT 1000-1200 BPS // In Proceedings of DSPA-2000

7. Моттль В.В., Мучник И.Б. Скрытые Марковские модели в структурном анализе сигналов. – М.: ФИЗМАТЛИТ, 1999

8. Steven Greenberg, Hannah Carvey, Leah Hitchcock, Shawn ChangTemporal properties of spontaneous speech–a syllable-centric perspective // Journal of Phonetics 2002, 31, pp.465-485.

9. Mermelstein P. (1976), Distance measures for speech recognition, psychological and instrumental in Pattern Recognition and Artificial Intelligence, C.H. Chen, Ed., pp. 374–388. Academic, New York.

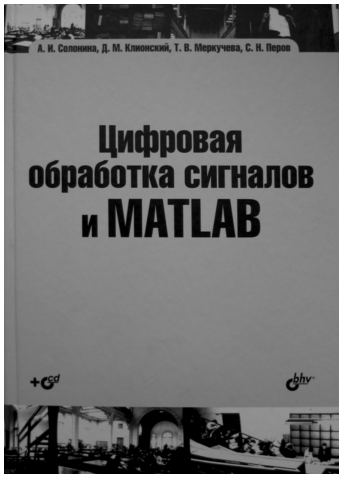
10. Davis S.B., and Mermelstein P. (1980), Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, in IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), pp. 357–366.

HYBRID VOICE ACTIVITY DETECTOR

Voznesenskya T.V., Kotov M.A., Lednov D.A.

This paper focuses on a hybrid voice activity detector (VAD) which is built from two detectors connected sequentially with different operating principles. The first detector, developed by the authors, is based pitch detection. This detector uses wavelet analysis applied in the spectral domain. It is shown that this detector is capable to separatespeechfrom the non-stationary background noise, which has a continuous spectrum, at low signal/noise ratio. The experimental results of detector investigation are shown. However, such a detector is not able to distinguish speech from the speech-like signals that have a line spectrum e.g., music. For this purpose a well-known detector is used, based on the likelihood ratio of statistic models of music and speech. The experimental results of hybrid VAD investigation are shown.

НОВЫЕ КНИГИ



Солонина А.И.

ЦИФРОВАЯ ОБРАБОТКА СИГНАЛОВ И MATLAB:
учеб. пособие / А.И. Солонина, Д.М. Клинский, Т.В. Меркучева, С.Н. Перов. – СПб.: БХВ-Петербург, 2013. – 512 с. (Учебная литература для вузов)
Москва: Техносфера, 2013. – 528 с.

Описываются базовые методы и алгоритмы цифровой обработки сигналов и средств их компьютерного моделирования в системе MATLAB. Даны основы алгоритмического языка MATLAB. Рассматриваются дискретные сигналы, линейные дискретные системы, дискретное преобразование Фурье с использованием алгоритмов БПФ, синтез и анализ КИХ- и БИХ-фильтров, в том числе с фиксированной точкой, спектральный анализ сигналов, многоскоростная обработка сигналов и адаптивная цифровая фильтрация.

Технология обучения в процессе компьютерного моделирования на основе созданных авторами программ или графического интерфейса пользователя MATLAB расширяет теоретические знания и позволяет понять многие важные проблемы и аспекты практического применения методов и алгоритмов ЦОС. На прилагаемом к книге CD хранятся обучающие программы и таблицы исходных данных.

Предназначена для студентов, аспирантов и преподавателей вузов, а также специалистов в области цифровой обработки сигналов.

У в а ж а е м ы е а в т о р ы!

Редакция научно-технического журнала «Цифровая обработка сигналов» просит Вас соблюдать следующие требования к материалам, направляемым на публикацию:

1) Требования к текстовым материалам и сопроводительным документам:

1. *Текст - текстовый редактор Microsoft Word.*
2. *Таблицы и рисунки должны быть пронумерованы. На все рисунки, таблицы и библиографические данные указываются ссылки в тексте статьи.*
3. *Объем статьи до 12 стр. (шрифт 12). Для заказных обзорных работ объем может быть увеличен до 20 стр.*
4. *Название статьи на русском и английском языках.*
5. *Рукопись статьи сопровождается:*
   - *краткой аннотацией на русском и английском языках;*
   - *номером УДК;*
   - *сведениями об авторах (Ф.И.О., организация, должность, ученая степень, телефоны, электронная почта);*
   - *ключевыми словами;*
   - *актом экспертизы (при наличии в вашей организации экспертной комиссии).*

2) Требования к иллюстрациям:

- *Векторные (схемы, графики) - желательно использование графических редакторов Adobe Illustrator или Corel DRAW.*
- *Растровые (фотографии, рисунки) - М 1:1, разрешение не менее 300dpi, формат tiff, jpg.*