Московский технический университет связи и информатики

ТЕХНИКА ЗВУКОВОГО ВЕЩАНИЯ

УДК 577.4:621.396

А.А. Кудинов

Автоматическое определение типа звучания

Автоматическое определение типа звучания относится к задачам создания искусственного интеллекта. По сути — это задача распознавания образов, и, следовательно, задача статистическая. Будем различать четыре типа звучания (класса распознаваемых образов) в сигналах звукового вещания:

1. музыка,
2. речь,
3. вокал (певческая речь, речь музыкально организованная),
4. вокал + музыка (пение с аккомпанементом).

Соответственно, задачу автоматического определения типа звучания будем понимать как задачу *отнесения* (т.е. принятия решения о принадлежности) *реализации сигнала* звукового вещания *к одному из* перечисленных классов. Одной из важнейших составляющих решения задачи распознавания образов является определения набора признаков, достаточного для достижения заданной достоверности распознавания.

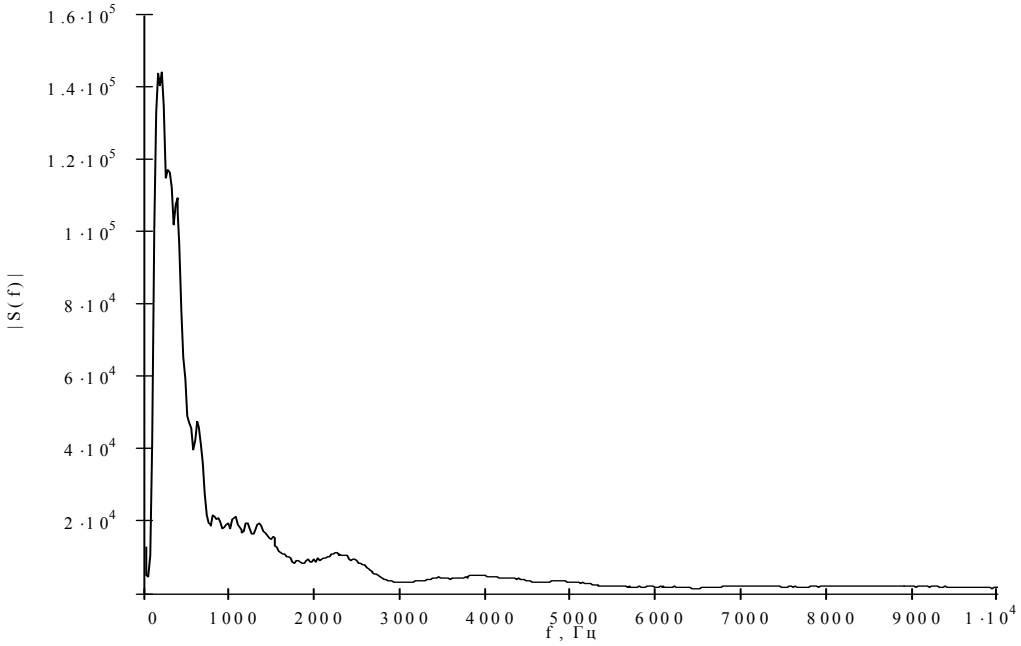
Музыкальный и речевой сигнал имеют различную природу: музыкальный сигнал отличается от всех звуковых сигналов определённой организацией – ритмической и гармонической, т.е. временной и спектральной. Спектральная организация происходит на двух уровнях:

1. тональные музыкальные звуки квазипериодичны, т.е. имеют дискретный спектр;
2. законы музыкальной гармонии требуют, чтобы звуки музыкального произведения принадлежали к одной тональности, т.е., если не учитывать незначительных отклонений, высоты тонов звуков принадлежат дискретному множеству.

Абсолютно точное выполнение этих двух условий означает, что частоты гармоник тональных музыкальных звуков также принадлежат дискретному множеству. Естественно, в реальных сигналах имеют место отклонения. Тем не менее, спектр речевого сигнала подобной организацией не обладает. Во-первых, высота тона то-

нальных звуков речи не подчиняется гармоническим законам, высота тона непостоянна даже на длительности фонемы, отклонения частот обертонов от значений, кратных ОТ, тональных звуков речи по относительным значениям превосходят относительные отклонения частот обертонов музыкальных звуков. Во-вторых, кроме тональных звуков (гласные), речевой сигнал содержит нетональные - звуки, соответствующие согласным. Спектры этих звуков не являются дискретными.

а)



б)

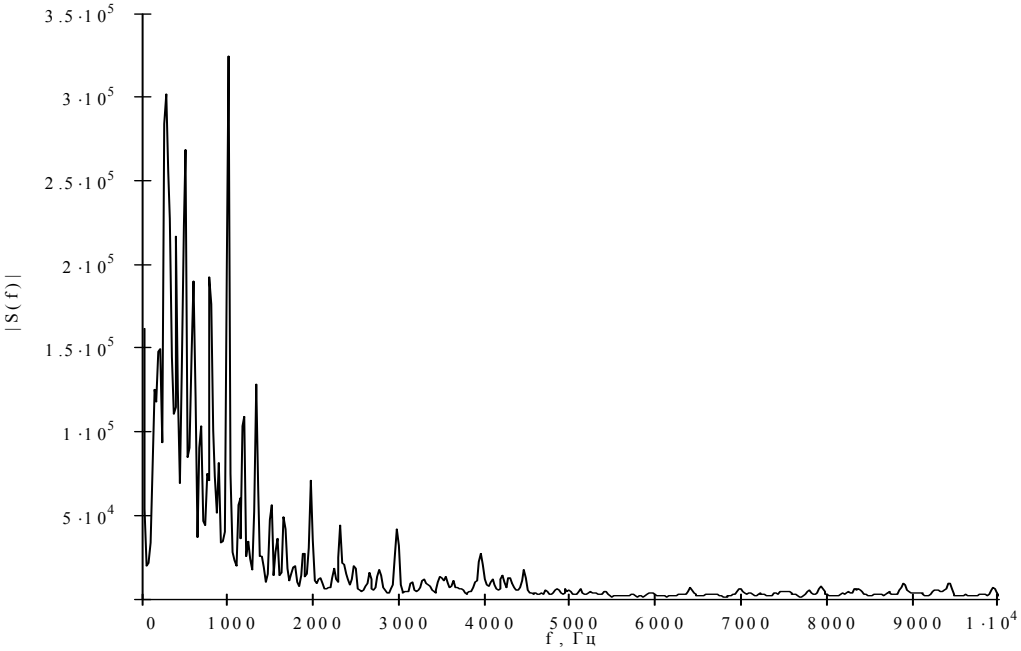
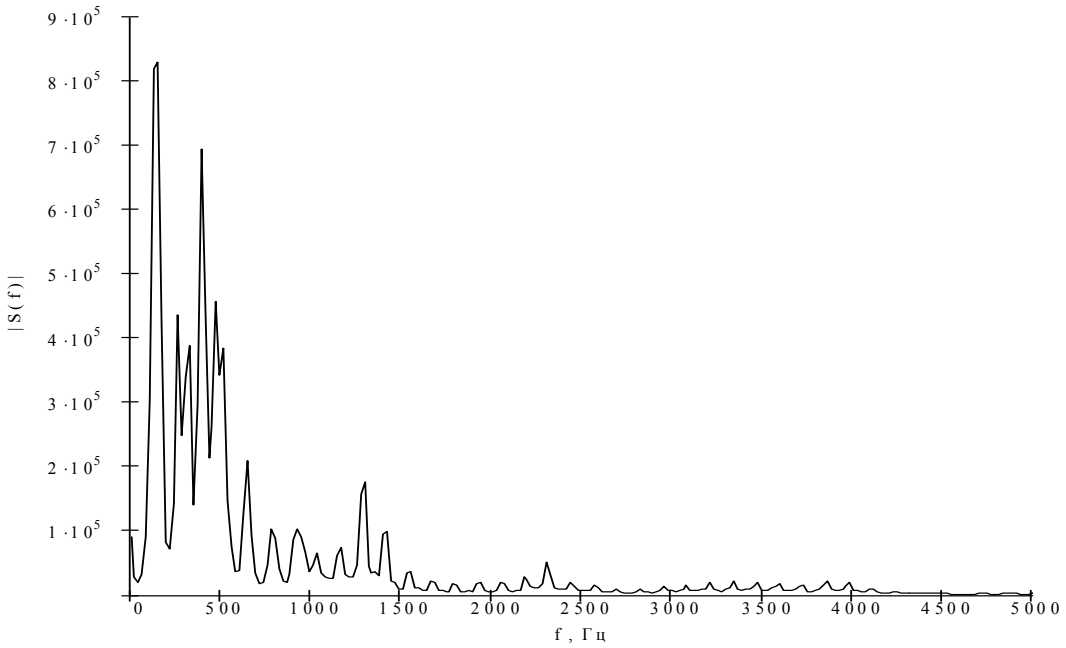


Рис. 1. Усреднённые амплитудные спектры сигналов: а) речевого и б) музыкального (длительность 2 минуты)

Принимая во внимание перечисленные различия свойств речевых и музыкальных сигналов, очевидным кажется применение в качестве признаков параметров амплитудного спектра. На рисунке 1 приведены усреднённые амплитудные спектры речевого и музыкального сигналов. Длительности сигналов, на которых проводилось усреднение больше одной минуты. Различия очевидны. Однако при длительности фрагментов 1 секунда принципиальных отличий нет – см. рисунок 2.

а)



б)

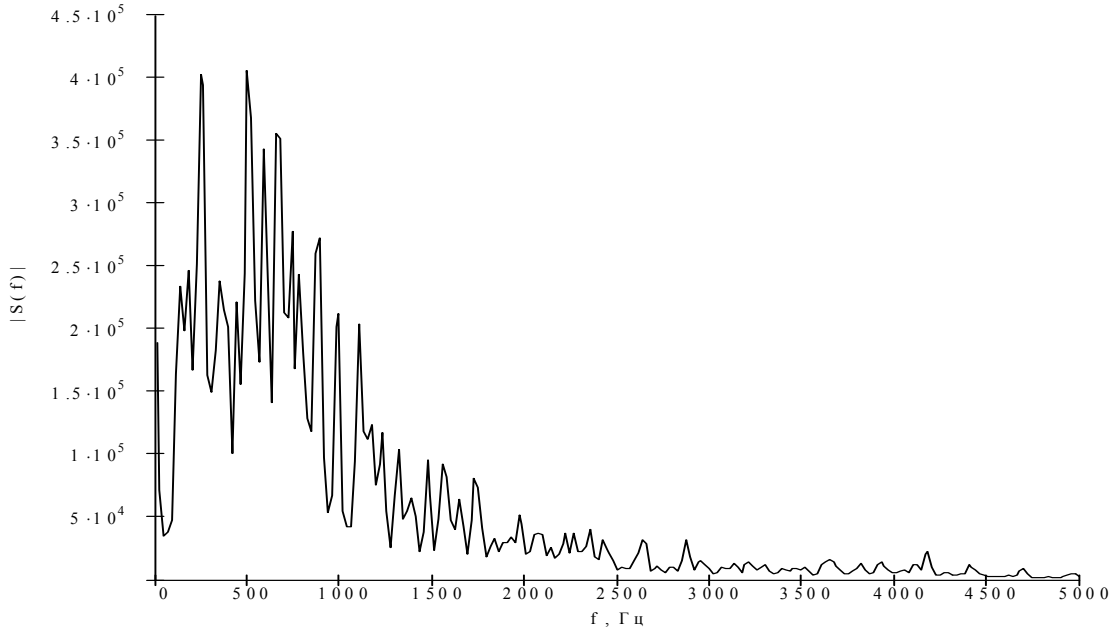
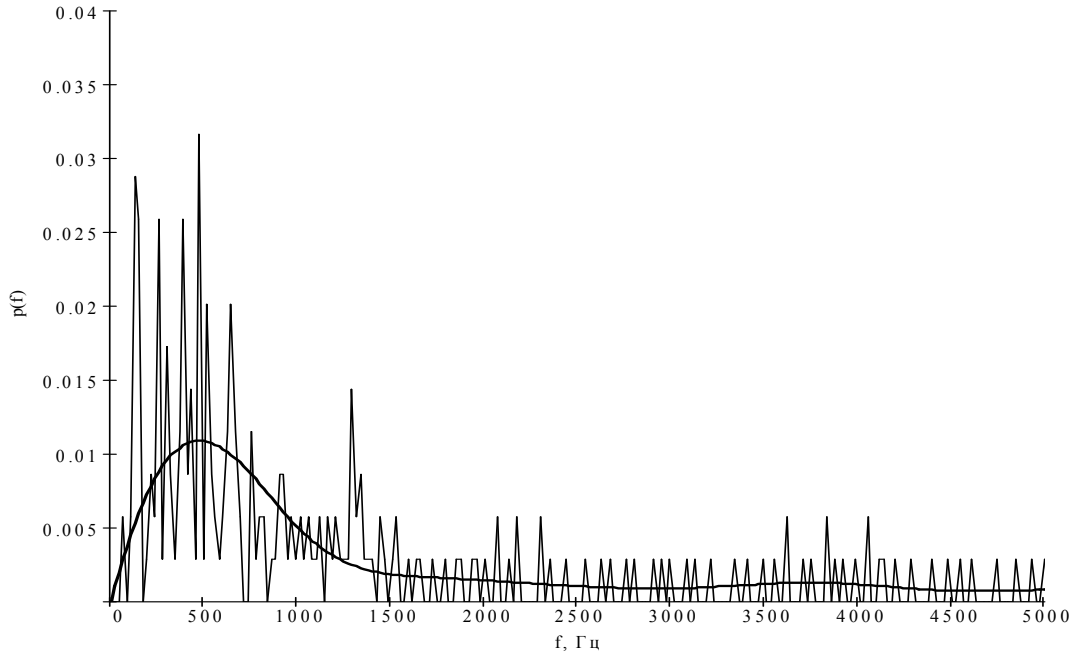


Рис. 2 Усреднённые амплитудные спектры сигналов: а) речевого и б) музыкального (длительность 1 секунда)

Почему так важна длительность, на которой производится усреднение амплитудного спектра? В устройствах обработки, учитывающих тип звучания, перенастройка параметров алгоритма обработки должна происходить как можно быстрее и как можно незаметнее на слух, т.е. должно быть максимально сокращено время определения типа звучания и минимизирована ошибка принятия решения.

а)



б)

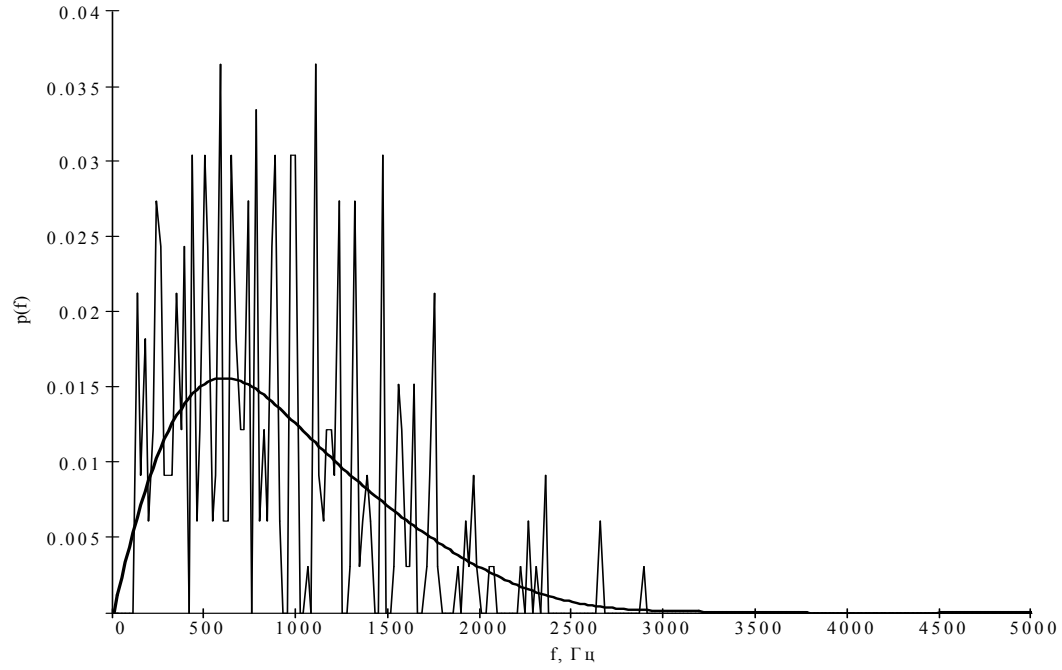


Рис. 3 Относительные частота появления максимумов в спектрах сигналов: а) речевого и б) музыкального (длительность 1 секунда)

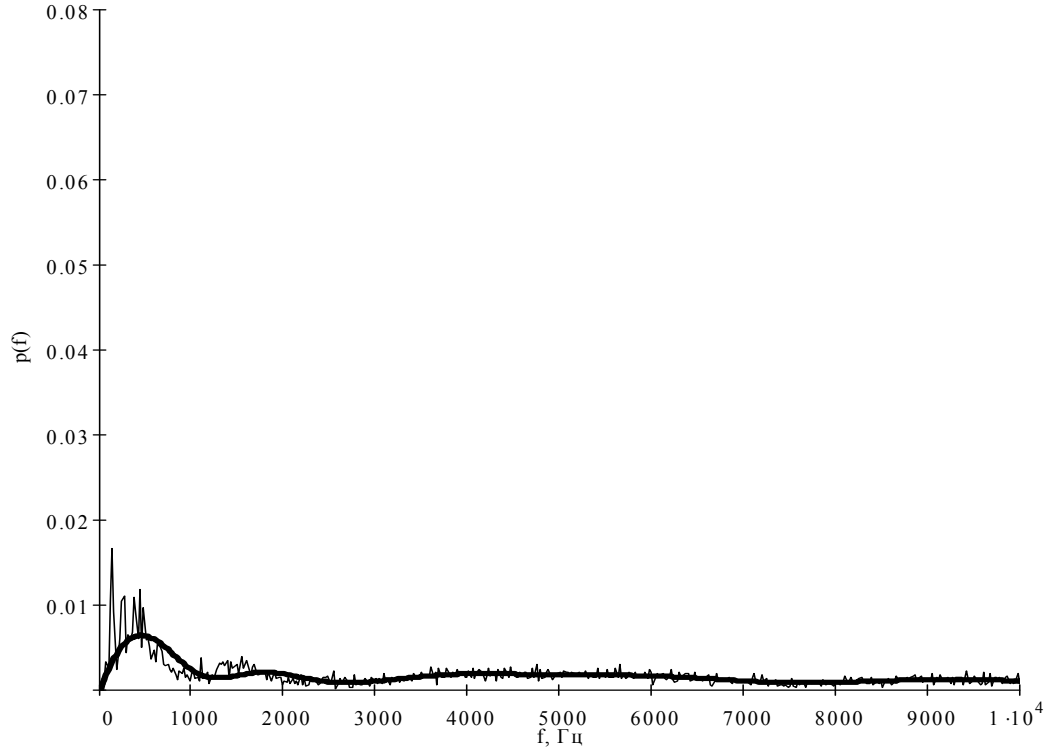
Для принятия решения о типе звучания по коротким (порядка 1 секунды) реализациям звукового сигнала предлагается использовать параметры статистики максимумов спектра. При обработке сигнала в цифровом виде спектр сигнала вычисляется путём дискретного преобразования Фурье. Т.е. получаемый амплитудный спектр – функция дискретного аргумента (частоты i-того члена ряда Фурье). Будем считать, что на частоте i-того члена ряда Фурье имеется максимум, если выполняются 2 условия:

1. амплитуда i-того члена больше амплитуд двух соседних членов: $A_{i-1} < A_i > A_{i+1}$
2. амплитуда i-того члена превышает некий порог $A_i > A_{i\,порог}$

Обсуждать более подробно процедуру выделения максимумов амплитудного спектра полагается неуместным. Подсчитав относительные частоты появления максимумов на частотах членов ряда Фурье, получим статистику максимумов спектра $p(f_i)$ - эмпирическую функцию распределения. На рисунке 3 приведены огибающие эмпирических функций распределения появления максимума спектра тех же фрагментов, чьи амплитудные спектры приведены на рисунке 2. Для наглядности графики построены в одном масштабе. Жирной линией на рисунках изображена сглаженная огибающая эмпирической функции распределения. Основное отличие таких распределений речевых и музыкальных сигналов состоит в том, что в музыкальных сигналах наиболее часто максимумы возникают на частотах до 3 кГц ÷ 6 кГц, в речевых сигналах $p(f_i)$ спадает медленнее. Из условия нормировки $\sum_i p(f_i) \equiv 1$ следует, что статистика частот максимумов спектра речевого сигнала более пологая функция, чем статистика музыкального сигнала — меньше максимальное значение, но распределение более равномерное. На рисунке 4 приведены эмпирические распределения для фрагментов длительностью 15 секунд. Для наглядности графики построены в одном масштабе.

Сглаженная функции распределения $Ps(f_i)$ (жирные линии на рисунках 3 и 4) в данном случае — сплайн интерполяция гистограммы. Количество значений эмпирической функции распределения 1024 (2048 отсчётов в выборке сигнала), количество интервалов гистограммы 1024 / 16 = 64. Распределения, полученные для речевых сигналов хорошо аппроксимируются таким образом, распределения, полученные для музыкальных сигналов — функции сильно флуктуирующие и аппроксимируются плохо – см. рисунок 5.

а)



б)

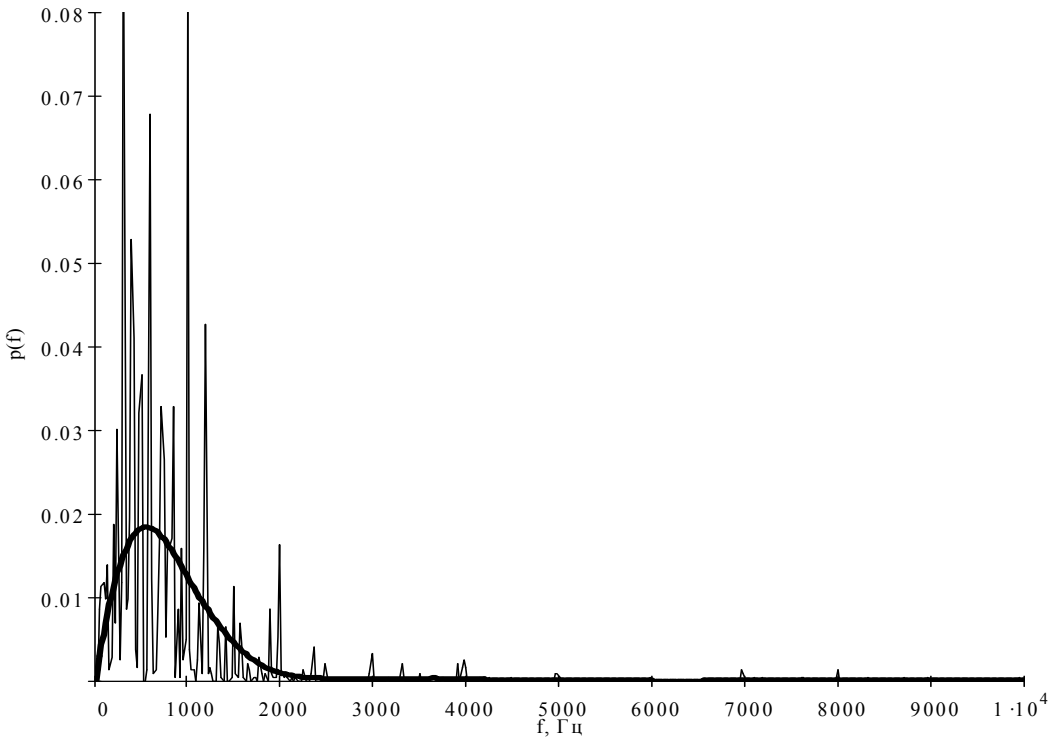
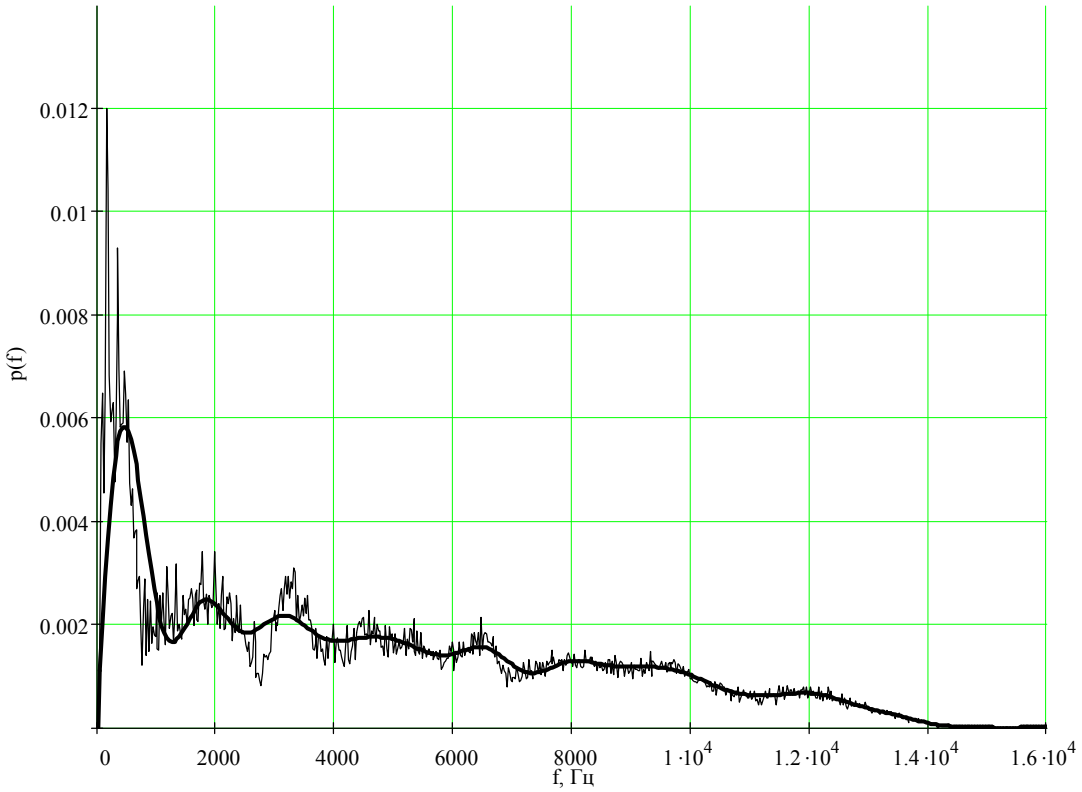


Рис. 4. Относительные частота появления максимумов в спектрах сигналов:
а) речевого и б) музыкального (длительность 15 секунд)

а)



б)

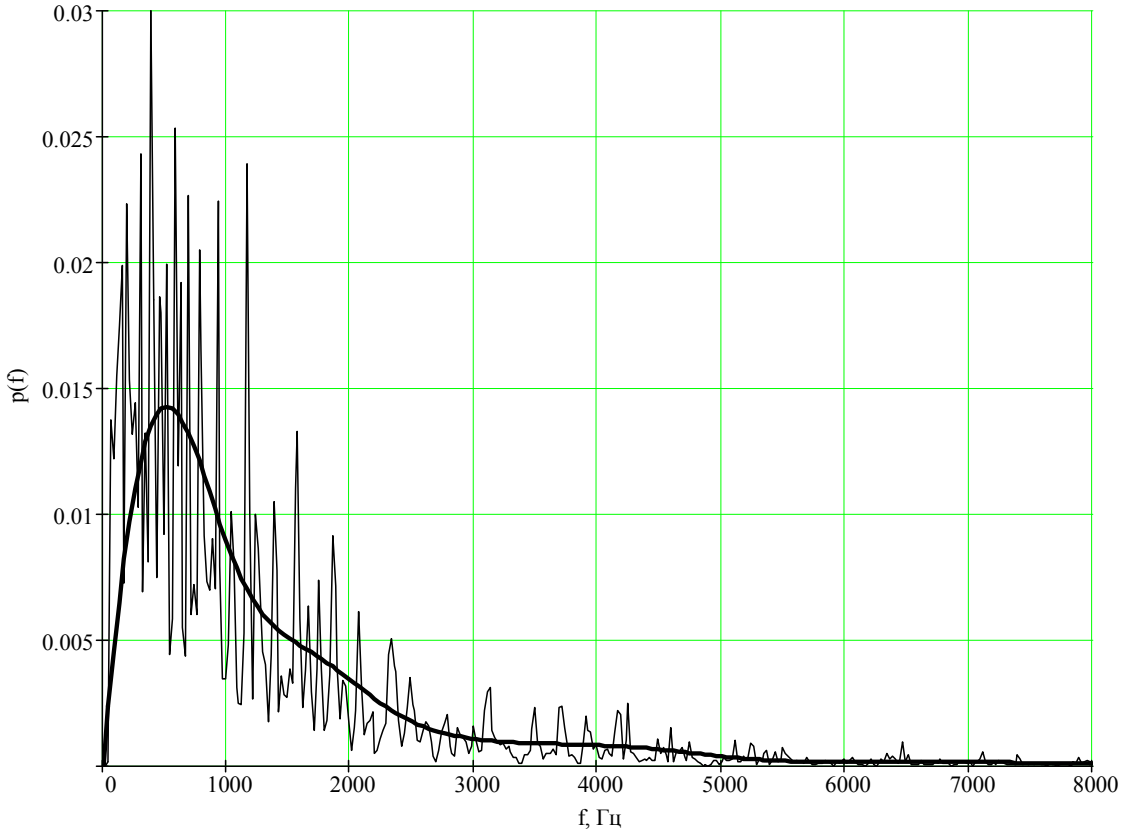


Рис. 5 Относительные частота появления максимумов в спектрах сигналов:
а) речевого и б) музыкального (длительность 2 минуты)

Очевидно, что при выборе одинакового способа аппроксимации отклонение сглаженной функции от оригинальной эмпирической функции распределения будет меньше для речевых сигналов. Таким образом, мы определились с выбором признаков, по которым будет выноситься решение о принадлежности реализации звукового сигнала к одному из типов – речь или музыка. Ограничимся двумя признаками:

1. Максимальное значение эмпирической функции распределения, обозначим его *M*

   $$M = \max_i [\, p(f_i)]$$

2. Среднеквадратическое отклонение сглаженной функции распределения от оригинальной, обозначим его *E*

   $$E = \frac{\sum_{i=0}^{N-1} \left| p(f_i) - Ps(f_i) \right|^2}{N}$$, где *N* — количество членов ряда Фурье с положительными частотами.

Автором были проанализированы 112 фрагментов звуковых сигналов различных типов звучания. Исходными были 16 фрагментов длительностью 2 минуты, из каждого было выделено ещё по 6 фрагментов длительностями соответственно 500 мс, 1 с, 2 с, 5 с, 15 с и 30 с. Из этих 16 фрагментов:

- 5 речевых
  английская речь (диктор мужчина),
  русская речь (диктор мужчина),
  женская речь (новости радиостанции «Маяк 24»),
  мужская речь (корреспондент-мужчина по телефону),
  мужская речь (звуковое сопровождение зарубежного фильма с русским переводом).
- 5 музыкальных
  современная инструментальная музыка (музыка Андрея Петрова),
  3 фрагмента оркестровых произведений (Г. Свиридов, А. Глазунов),
  рок-музыка.
- 4 вокальных
  мужской вокал - тенор (Лучано Паваротти),
  женский вокал – сопрано (Кэтрин Ботт),
  мужской вокал – рок-певец (Стив Оверлэнд),
  мужской квартет «Тризна».

- 2 фрагмента – вокал с аккомпанементом
  смешанный хор с симфоническим оркестром,
  современная акустическая музыка.

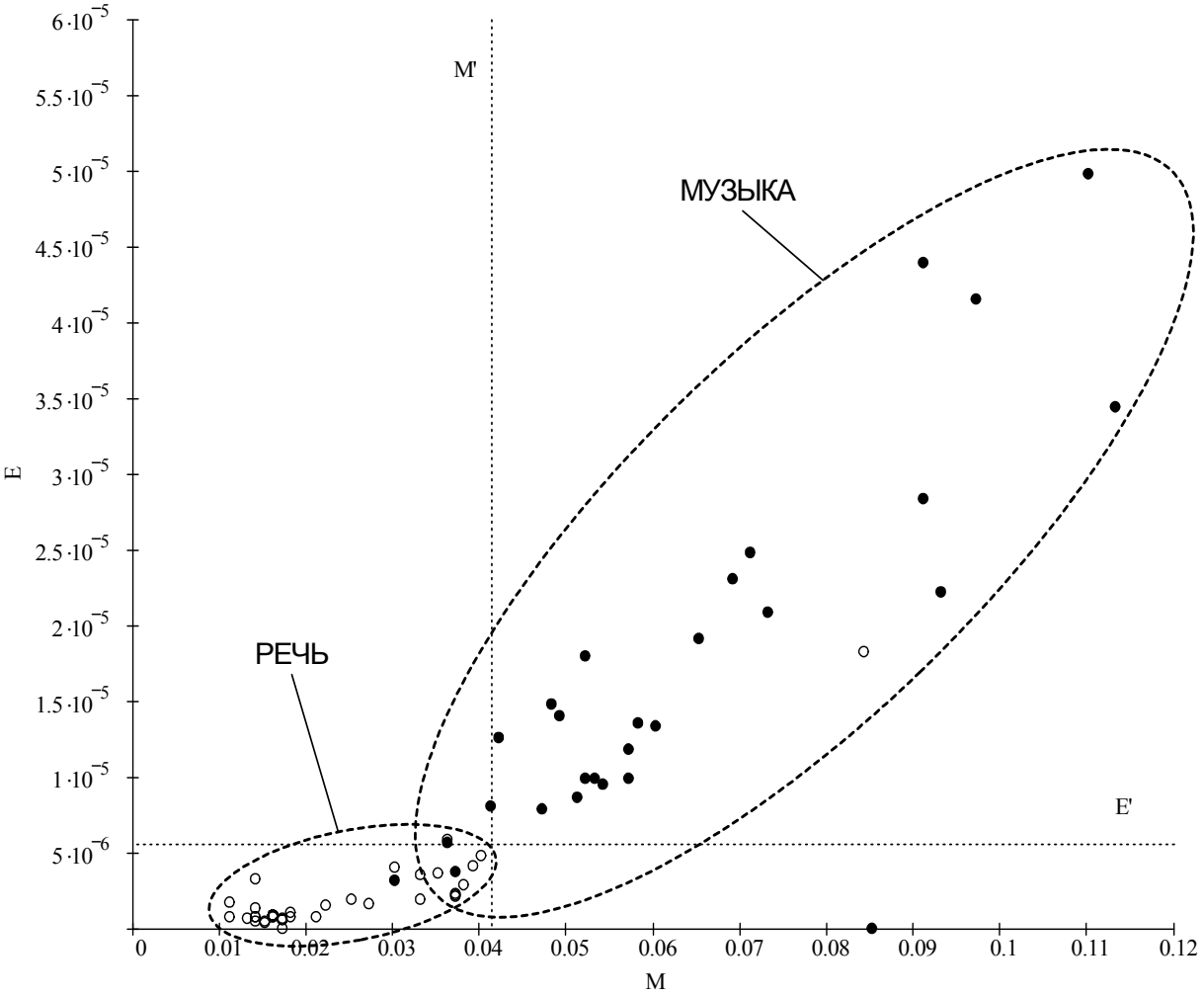


Рис. 6. Поле признаков для определения типа звучания – экспериментальные данные

На рисунке 6 приведено поле значений признаков для фрагментов различных длительностей речевых и музыкальных фрагментов (всего 70 фрагментов). Точки соответствуют музыкальным фрагментам, кружки — речевым. Видно, что области значений признаков речевых и музыкальных фрагментов пересекаются, хотя и незначительно (учитывая то, что нанесены точки, соответствующие всем длительностям, в том числе коротким – 0,5 с и 1 с). Пунктирными линиями определены примерные границы значений признаков *E'* и *M'* для вынесения решения о принадлежности к типу звучания. Определение точных границ выполняется путём минимизации вероятности ошибочного решения по более полной статистике значений признаков. На поле признаков сознательно не изображены точки, соответствующие двум оставшимся типам звучания – вокалу и вокалу с аккомпанементом. Эти точки занимают промежуточное значение между областями речи и музыки и для наглядности

выраженности этих двух областей не изображены.

Данный материал не представляет собой описание алгоритма или способа решения задачи автоматического определения типа звучания. Предлагается возможной путь решения этой задачи. Возможно, есть более информативные параметры сигнала, чем усреднённый амплитудный спектр или статистика максимумов спектра, или двух признаков для достаточно достоверного распознавания недостаточно. Но ясно, что задача автоматического определения типа звучания является статистической и вероятность вынесения правильного решения пропорциональна длительности анализируемого фрагмента, при этом существуют пути решения задачи.