Первичный анализ речевых сигналов.

Прежде чем рассказать о последних достижениях в теории распознавания речи, первую лекцию мы хотели бы посвятить первичному анализу речевых сигналов. Начнем мы с известного метода Мел-кепстральных коэффициентов, а также расскажем об операторах рассеивания, которые строятся с помощью вейвлетов.

1 Мел-кепстральные коэффициенты (MFCC).

Первый шаг в анализе речевых данных – это выделение признаков, которые являются "хорошими" для идентификации лингвистического содержания и отбрасыванием всех остальных признаков, отвечающих за шум и эмоции.

Главное, что нужно понять о речи, это то, что звуки, воспроизводимые человеком, определяются формой голосового тракта, включая язык, зубы и т. д. Если мы сможем точно определить форму голосового тракта, то мы будем иметь точное представление о производимой фонеме. Форма голосового тракта описывается огибающей спектра, и задача MFCC состоит в том, чтобы точно представить эту огибающую.

Мел-кепстральные коэффициенты были введены S. Davis и P. Mermelstein ([1]). До этого основными характеристиками для распознавания речи были линейные коэффициенты предсказания и линейные кепстральные коэффициенты предсказания.

Шкала Мел соотносит воспринимаемую частоту или высоту чистого тона (мел) с фактической измеренной частотой (Гц). Люди гораздо лучше различают небольшие изменения высоты звука на низких частотах, чем на высоких. Эта зависимость не совсем линейная и описывается следующей формулой:

$$M(f) = 1127,01048 \ln(1 + f/700). \tag{1}$$

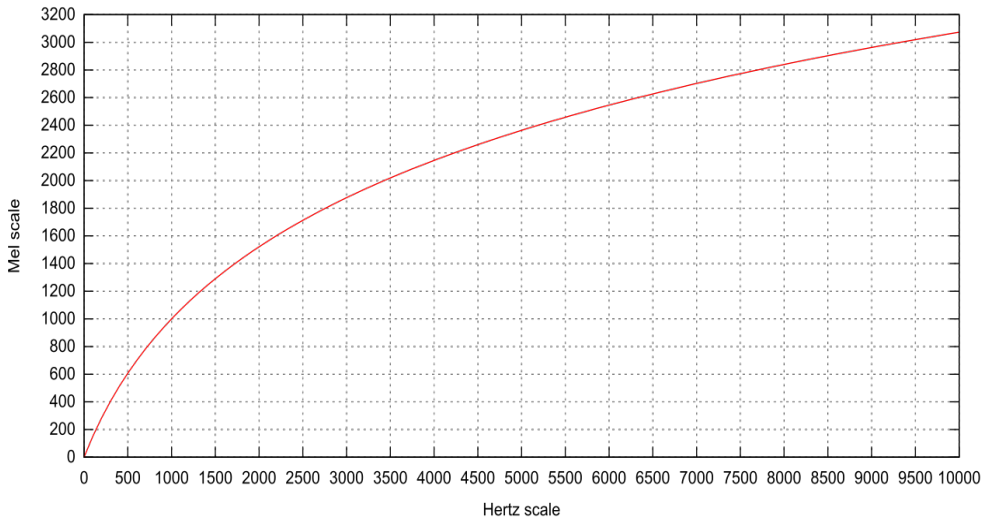

Рис. 1: График зависимости частоты от мел.

Обратное преобразование из мел в частоту выглядит следующим образом:

$$M^{-1}(m) = 700(e^{m/1127,01048} - 1). \tag{2}$$

Вычисление мел-частотных кепстральных коэффициентов включает в себя следующие шаги:

1. Необходимо разделить исходный сигнал на фреймы. Размер фрейма обычно выбирается от 20 до 40 мс, так как считается, что речевой сигнал на этом промежутке не сильно меняется.

Речевой сигнал записываем в виде:

$$x(n),\ 0 \le n < N, \text{где } N \text{ – размер фрейма или длина окна,}$$

$$x_j(n) \text{– j-ый фрейм.}$$

Следующие шаги применяются для каждого отдельного фрейма.

2. Речевой сигнал конечен и не является периодическим, поэтому из-за разрывов на его концах при применении преобразования Фурье проявляется эффект утечки. Для того, чтобы снизить его влияние на результат, каждый кадр умножается на оконную функцию Хемминга:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right),\ 0 \le n \le N-1.$$

К получившемуся результату применяем дискретное преобразование Фурье:

$$X_j(k) = \sum_{n=0}^{N-1} x_j(n)w(n)e^{-\frac{2\pi i}{N}kn},\ 0 \le k < N,$$

где j– номер фрейма.

3. Вычисляем периодограмму для каждого фрейма по следующей формуле:

$$P_j(k) = \frac{|X_j(k)|^2}{N}.$$

4. Вычисляем блок мел-фильтров. Для этого треугольные фильтры (от 20 до 40) умножаются на периодограмму и суммируются. В результате мы получим энергии набора фильтров.

Каждый треугольный фильтр моделируются с помощью следующей функции:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \le k < f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases},$$

где m – это число фильтров, которое мы хотим получить. Зная число фильтров (обычно 26) и диапазон интересующих нас частот, функции $f(\cdot)$ можно найти, используя формулы (1) и (2).

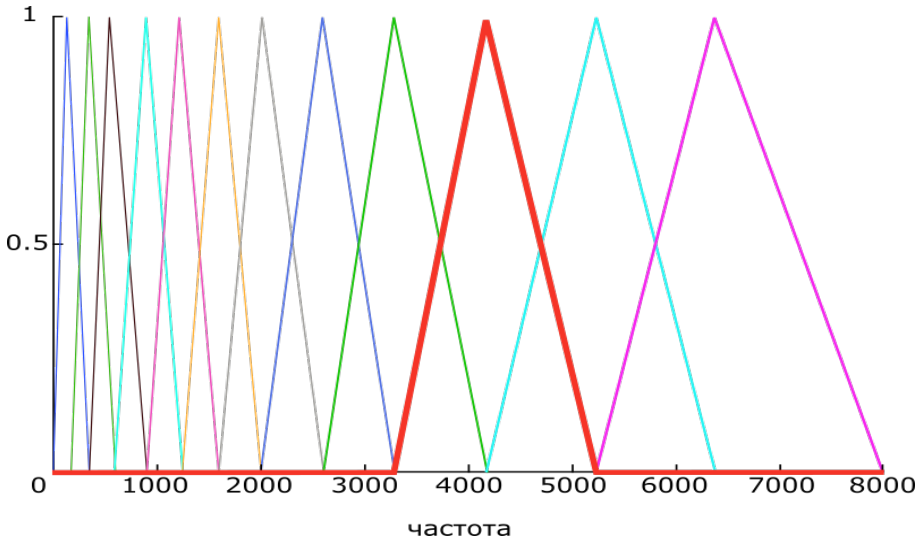

Рис. 2: Фильтры собираются в области низких частот, обеспечивая более высокое "разрешение" там, где это необходимо для распознавания.

5. Полученные энергии логарифмируются. Это также мотивируется человеческим слухом: мы не слышим громкость в линейном масштабе. Обычно, чтобы удвоить воспринимаемую громкость звука, нам нужно затратить в 8 раз больше энергии. Это означает, что большие колебания энергии могут звучать не так уж и по-другому, если звук с самого начала громкий. Эта операция сжатия делает наши функции более близкими к тому, что на самом деле слышат люди.

Мы получаем некоторый набор коэффициентов, которые еще не являются MFCC:

$$S_j(m) = \ln \sum_{k=0}^{N-1} P_j(k) H_m(k),\ 0 \le m < M.$$

6. Далее, используя дискретное косинусное преобразование, получим **мел-кепстральные** коэффициенты:

$$c_j(n) = \sum_{m=0}^{M-1} S_j(m) \cos(\pi n(m+1/2)/M),\ 0 \le n < M.$$

Причина, по которой мы делаем дискретное косинусное преобразование в следующем. Наши фильтры пересекаются, а энергии фильтров достаточно коррелируют. Дискретное косинусное преобразование декоррелируют их. Но обратите внимание, только 12 из 20 коэффициентов сохраняются. Это связано с тем, что более высокие коэффициенты представляют быстрые изменения энергий набора фильтров, и оказывается, что эти быстрые изменения фактически ухудшают производительность рапознавания речи, поэтому мы получаем небольшое улучшение, отбрасывая их.

2 Операторы рассеивания. [2]

Мелкепстральные коэффициенты – эффективное описание аудио сигнала для измерения спектральной энергии на коротких временных окнах длиной в 23 мс. Однако эти измерения теряют нестационарную спектральную информацию. Покажем, что эта потерянная информация может быть изучена с помощью преобразования рассеивания.

Рассмотрим вейвлет фильтр банк, состоящий и функций $\{\psi_j\}_{j<J+P}$, сконструированных с помощью растяжения базового фильтра ψ, у которого ширина октавной полосы $1/Q$:

$$\psi_j(t) = a^{-j}\psi(a^{-j}t),\ \text{где } a = 2^{1/Q} \text{ и } j < J. \tag{3}$$

Эти фильтры можно рассматривать, как вейвлеты растяжения. Фильтр ψ нормирован так, что его носителем во временной области является 1 секунда, а в частотной области носитель покрывает интервал

$$[2Q\pi - \pi, 2Q\pi + \pi].$$

Для $j < J$ во временной области носитель ψ_j из (3) содержится в интервале $[0, a^J]$, а в частотной области носитель покрывает интервал

$$[2Q\pi a^{-j} - \pi a^{-j}, 2Q\pi a^{-j} + \pi a^{-j}].$$

Это следует из того, что коэффициенты Фурье $\hat{\psi}_j(\omega) = \hat{\psi}(a^j\omega)$. Для частот ниже, чем $2\pi Q a^{-J}$, фильтр банк состоит из P фильтров ($J \le j < P + J$), имеющие ширину частотной полосы $2\pi a^{-J}$, как и у ψ_J, и временной носитель $[0, a^J]$. Хотя эти низкочастотные фильтры не являются растяжением ψ, для простоты будем называть их вейвлетами.

Определим оператор с помощью описанных вейвлетов:

$$W_J x(t) = \begin{pmatrix} x * \phi_J(t) \\ x * \psi_j(t) \end{pmatrix},\ \ j < P + J, \tag{4}$$

где символ $*$ обозначает свертку. Первый фильтр ϕ_J – это низкочастотный фильтр, который считается для частоты, ниже самой низкой частоты ψ_j. Его носитель во временной области $[0, a^J]$. В частотной области этот фильтр покрывает интервал $[-\pi a^{-J}, \pi a^{-J}]$. Этот интервал не покрывается другими вейвлет-фильтрами. Предполагается, что все фильтры удовлетворяют условию:

$$\overline{f}(-x) = f(x), \tag{5}$$

где черта обозначает сопряжение. Вейвлет фильтры построены так, чтобы для всех частот ω выполнялось неравенство

$$1 - \varepsilon \le |\hat{\phi}_J(\omega)|^2 + \frac{1}{2} \sum_{j<P+J} |\hat{\psi}_j(\omega)|^2 + |\hat{\psi}_j(-\omega)|^2 \le 1 \tag{6}$$

для малых ε. Используя это неравенство можно показать, что вейвлет-преобразование W_J является сжимающим, а для $\varepsilon = 0$ унитарным. Откуда следует, что x может быть восстановлено из вейвлет-преобразования.

В численных приложениях мы используем фильтры Габора $\psi(t) = \theta(t)e^{i2\pi Qt}$, которые удовлетворяет (5) и (6) для $\varepsilon = 0.02$. Здесь $\theta(t)$ – гауссиан, $Q = 16$ и $P = 23$, J находится из соотношения $a^J = T$, где $T = 800$ мс. Получившийся фильтр банк изображен на следующем рисунке.

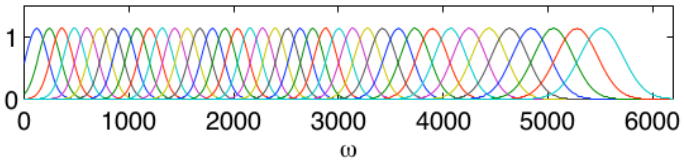

Рис. 3: Вейвлет-преобразование, состоящее из фильтров Габора в сэмплированной частоте 1125 Hz.

2.1 Вейвлеты рассеивания.

Выберем максимальный вейвлет масштаб $a^J = T$. Рассмотрим выражение

$$|x * \psi_j| * \phi_J(t),$$

которое измеряет сигнальную амплитуду x в частотном интервале, покрытым ψ_j, и усредненное в окрестности t длиной $T = a^J$. Чем больше T, тем большие информации теряется при усреднении.

Для восстановления информации, утерянной при усреднении, вычислим то, что называется **коэффициентами второго порядка**. Для этого рассмотрим $|x * \psi_{j_1}|$, как новый сигнал для любого $j_1 < J + P$, и вычислим $|x * \psi_{j_1}| * \psi_{j_2}(t)$ для любого $j_1, j_2 < J + P$. Заметим, что $|x * \psi_{j_1}| * \phi_J(t)$ может быть записан, как низкочастотный компонент вейвлет-преобразования от $|x * \psi_{j_1}|$, т.е.

$$W_J|x * \psi_{j_1}|(t) = \begin{pmatrix} |x * \psi_{j_1}| * \phi_J(t) \\ |x * \psi_{j_1}| * \psi_{j_2}(t) \end{pmatrix}_{j_2 < P+J}.$$

Снова усредним полученное выражение с помощью низкочастотных фильтров:

$$||x * \psi_{j_1}| * \psi_{j_2}(t)| * \phi_J(t). \tag{7}$$

Значение этих коэффициентов дает дополнительную информацию в масштабах a^{j_1} и a^{j_2}. Коэффициенты (7) называются **коэффициентами рассеивания**, так как они вычисляют взаимодействие сигнала с двумя вейвлетами ψ_{j_1} и ψ_{j_2}. Они измеряют амплитуду временных изменений $|x * \psi_{j_1}(t)|$ в частотных интервалах, покрытых вейвлетами ψ_{j_2}. Для сигнала длины N существует $Q \log_2 N/Q$ коэффициентов первого порядка и $Q^2/2 \log_2^2(N/Q^2)$ коэффициентов второго порядка.

На следующем рисунке показаны коэффициенты рассеивания первого порядка для музыкальной записи, сделанной при 11025 Hg, вычисленных для $T = 800$ мс.

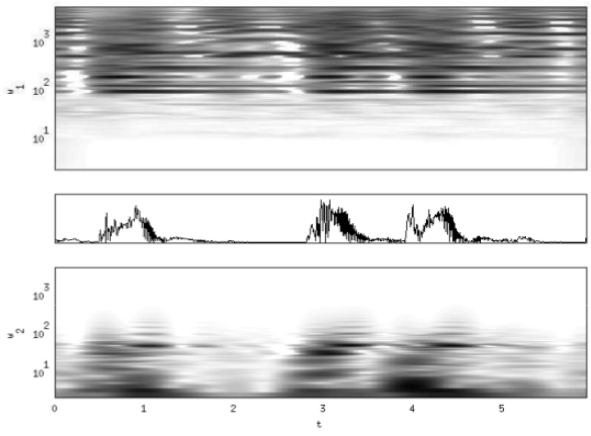

Рис. 4: Вверху: $\log(|x * \psi_{j_1}| * \phi_J(t))$ функция времени t и частоты $\omega_1 = 2\pi Q a^{-j_1}$ для $T = a^J = 800$ мс. В середине график $|x * \psi_{j_1}|$ для $\omega_1 = 855$ Hz. Внизу: $\log(||x * \psi_{j_1}| * \psi_{j_2}| * \phi_J(t))$ как функция от t и $\omega_2 = 2\pi Q a^{-j_2}$ для фиксированного масштаба a^{j_1} для $|x * \psi_{j_1}|$, показанной выше.

Усреднение в (7) снова влечет потерю высоких частот, которые могут быть восстановлены с помощью нового вейвлет-преобразования. Таким образом, мы получаем итерацию данной процедуры.

Пусть U_J – вейвлет-модульный оператор, который вычисляет модуль комплексных вейвлет-коэффициентов, сохраняя фазу $x * \phi_J$:

$$U_J x(t) = \begin{pmatrix} x * \phi_J(t) \\ |x * \psi_j(t)| \end{pmatrix}_{j<P+J}$$

На первом шаге **преобразования рассеивания** применяется оператор $U_J x$, вычисляя низкочастотный сигнал $x * \phi_J(t)$ и модуль $|x * \psi_{j_1}|$. На следующем шаге каждое $|x * \psi_{j_1}|$ трансформируется с помощью U_J в $|x * \psi_{j_1}| * \phi_J$ и вычисляет $||x * \psi_{j_1}| * \psi_{j_2}|$. Эти коэффициенты снова трансформируются с помощью U_J и так далее.

Применяя такое преобразование m раз и отбрасывая коэффициенты, нефильтрованные ϕ_J, мы имеем вектор рассеивания порядка $m + 1$:

$$S_J x(t) = \begin{pmatrix} x * \phi_J(t) \\ |x * \psi_{j_1}| * \phi_J(t) \\ |x * \psi_{j_1}| * \psi_{j_2}| * \phi_J(t) \\ \dots \\ ||\dots|x * \psi_{j_1}|\dots| * \psi_{j_m}| * \phi_J(t) \end{pmatrix}_{j_1, j_2, \dots j_m < P+J} . \tag{8}$$

Для $m = 0$, $S_J x(t) = x * \phi_J(t)$. Норма оператора S_J вычисляется следующим образом:

$$\|S_J x\|^2 = \sum_m \sum_{j_1,\dots,j_m<J+P} \|||x * \psi_{j_1}| \cdots * \psi_{j_m}| * \phi_J\|^2.$$

Можно показать, что оператор S_J является сжимающим и унитарным при условии, что вейвлет-фильтры удовлетворяют (6) с $\varepsilon = 0$. В работе [3] показано, что для вейвлетов, удовлетворяющие некоторому критерию, энергия коэффициентов рассеивания порядка m стремится к нулю, когда $m \to \infty$. В следующей таблице показаны средние значения $\|S_J x\|/\|x\|$ по всем аудио сигналам в базе GTZAN. Заметим, коэффициенты

T	$m = 0$	$m = 1$	$m = 2$	$m = 3$
23 ms	23.7%	98.9%	99.6%	99.6%
93 ms	1.9%	97.7%	99.4%	99.4%
370 ms	1.2%	92.7%	99.3%	99.4%
1.5 s	1.0%	82.0%	98.9%	99.3%
5.9 s	0.99%	73.0%	98.1%	99.1%
22 s	0.97%	67.5%	96.5%	99.0%

Рис. 5: Среднее значение $\|S_J x\|/\|x\|$ по всем аудио сигналам x из базы GTZAN, просэмплированные при 11025 Hz, как функции от m и T.

первого и второго порядка сохраняют большую часть энергии сигнала в окне размера T. Поэтому далее вычислим оператор преобразования рассеивания порядка $m = 2$.

Сигналы $|x * \psi_{j_1}| * \phi_J(t)$ и $||x * \psi_{j_1}| * \psi_{j_2}| * \phi_J(t)$ равномерно сэмплируются на интервалах размера $T = a^J$. Оператор рассеивания второго порядка

$$S_J x(na^J) = \begin{pmatrix} x * \phi_J(na^J) \\ |x * \psi_{j_1}| * \phi_J(na^J) \\ |x * \psi_{j_1}| * \psi_{j_2}| * \phi_J(na^J) \end{pmatrix}_{j_1, j_2 < P+J} .$$

В работе [2] показано, что если $j_2 < j_1 + \log_a Q/2$, то $||x * \psi_{j_1}| * \psi_{j_2}| * \phi_J(t) \approx 0$. Поэтому коэффициенты второго порядка вычисляются только для $j_2 \geq j_1 + \log_a Q/2$ с помощью следующего алгоритма:

for $j_1 < P + J - 1$ **do**
    Compute $|x * \psi_{j_1}(a^{j_1} n)|$ $\forall n$
    Output $|x * \psi_{j_1}| * \phi_J(a^J n)$ $\forall n$
    for $j_2 = j_1 + \log_a(Q/2)$ to $P + J - 1$ **do**
        Compute and output $||x * \psi_{j_1}| * \psi_{j_2}| * \phi_J(a^J n)$ $\forall n$
    end for
end for

Замечание 1. *Обратимость оператор преобразования рассеивания сводится к обратимости оператора U_J из (2.1). Даже если доказать, что U_J – обратим (это можно сделать только для подходящих вейвлетов), x не может быть восстановлен точно из $S_J x$ конечного порядка m, так как все коэффициенты рассеивания порядка больше m являются множеством меры нуль. Для $T \leq 100$ ms большая часть сигнальной энергии сконцентрированная в коэффициентах первого порядка Рис.6. При увеличении T реконструкция x из коэффициентов первого порядка теряет большую часть информации, а при $T \geq 3s$ теряются все мелодические структуры. Второго порядка коэффициенты восстанавливают эту временную информацию при $T = 3s$.*

2.2 Косинусное логарифмическое рассеивание.

Вычисление косинусного логарифмического рассеивания похоже на вычисление мел-кепстральных коэффициентов из мел-фильтра.

Известно ([5], [4]), что многие музыкальные и голосовые звуки могут быть аппроксимированны с помощью модели "источник-фильтр" , в которой сигнал $x(t)$ можно представить с помощью возбуждения $e(t)$, отфильтрованного с помощью резонатора, соответствующего фильтру $h(t)$, т.е.

$$x(t) = e * h(t).$$

Аналогично мелкепстральным коэффициентам, компоненты фильтра h и источника e могут быть линейно разделены с помощью косинусного дискретного преобразования (DCT).

Фильтр $h(t)$ обычно такой, что $\hat{h}(\omega)$ – регулярная функция по ω. Предполагая, что $\hat{h}(\omega)$ – почти константа на частотном носителе функции $\hat{\psi}_{j_1}$, можно проверить, что

$$x * \psi_{j_1}(t) \approx \hat{h}(2\pi Q a^{-j_1}) \cdot e * \psi_{j_1}(t).$$

Откуда следует, что

$$\log |x * \psi_{j_1}| * \phi_J(t) \approx \log |\hat{h}(2\pi Q a^{-j_1})| + \log(|e * \psi_{j_1}(t)| * \phi_J(t)). \tag{9}$$

Так как $|\hat{h}(\omega)|$ – регулярная функция по ω, $\log |\hat{h}(2\pi Q a^{-j_1})|$ – регулярная функция по j_1, а $|e * \psi_{j_1}(t)|$ не является регулярной. Поэтому их нужно разделить с помощью DCT по j_1.

Аналогично (9), получаем

$$||x * \psi_{j_1}| * \psi_{j_2}| * \phi_J(t) \approx |\hat{h}(2\pi Q a^{-j_1})| \cdot ||e * \psi_{j_1}| * \psi_{j_2}| * \phi_J(t),$$

и следовательно,

$$\log(||x * \psi_{j_1}| * \psi_{j_2}| * \phi_J(t)) \approx \log |\hat{h}(2\pi Q a^{-j_1})| + \log(||e * \psi_{j_1}| * \psi_{j_2}| * \phi_J(t)). \tag{10}$$

Последнее выражение преобразуется с помощью DCT вдоль j_2, а затем вдоль j_1, что дает представление, параметризованное k_2 и k_1 соответственно.

Следующий рисунок показывает, как DCT эффективно декоррелирует логарифмические коэффициенты рассеивания и концентрирует энергию в нескольких первых коэффициентах.

Замечание 2. *В отличии от MFCC, которое зависило только от вектора параметров, преобразование рассеивания второго порядка зависит от двух параметров k_2 и k_1, т.е. эта целая матрица параметров. Первое слагаемое в (10), зависящее только от j_1, только вносит вклад в нулевой коэффициент DCT ($k_2 = 0$) вдоль j_2. Второе DCT разделяет низкочастотные компоненты вдоль j_1 от высокочастотных.*

Для классификации музыкального жанра в окончательном представлении используются косинусные логарифмические коэффициенты рассеивания (CLS), полученные с помощью сохранения только низкочастотных DCT коэффициентов, как и в случае MFCC. Для $m = 1$ сохраняются первые a_1 коэффициентов. Для $m = 2$ выбирается квадрат $k_1 < a_1$, $k_2 < a_2$.

2.3 Классификация.

Сравним результаты MFCC, Delta-MFCC и косинусного логарифмического преобразования рассеивания для музыкальной классификации на базе GTZAN. Эта база включает 10 жанров, каждый из которых содержит 100 клипов по 30 секунд каждый.

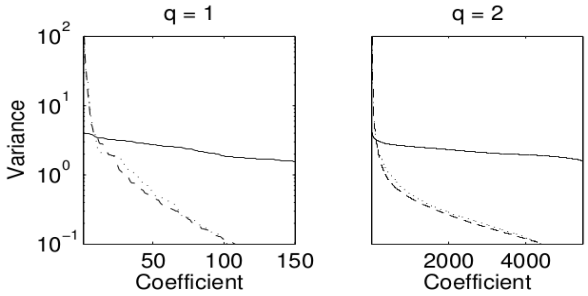

Рис. 6: Дисперсия логарифмических коэффициентов рассеивания для $m = 1$ и $m = 2$, вычисленная на базе GTZAN для $T = 1.5s$. Сплошая линия – это дисперсия логарифмических коэффициентов рассеивания. Пунктирная – дисперсия PCA базиса (Метод главных компонент), вычисленная на логарифмических коэффициентах рассеивания. Точечная кривая – дисперсия косинусных логарифмических коэффициентов рассеивания.

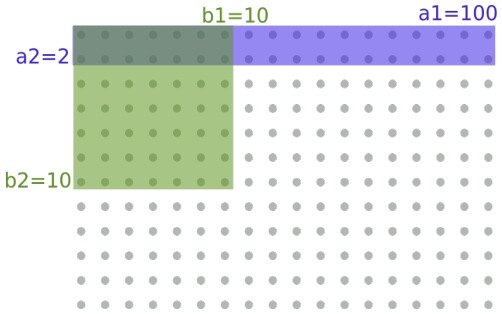

Рис. 7: Для численных результатов, представленных в статье, выбирается $a_1 = 100$, $b_1 = 10$, $a_2 = 2$, $b_2 = 10$. Размер представления меньше, чем 100 коэффициентов для $m = 1$ и 380 коэффициентов для $m = 2$.

Delta-MFCC коэффициенты определяются, как разница между MFCC коэффициентами двух последовательных фреймов и, таким образом, покрывают временной интервал двойного размера. Этот метод дополняет MFCC, обеспечивая информацию о временной аудио динамики на длинных временных интервалах.

Каждый аудио трек разлагался на фреймы длительностью T, которые были представлены с помощью MFCC, Delta-MFCC и CLS. Результаты представлены в следующей таблице.

T/classifier	0.023 s/PCA	0.19 s/PCA	1.5 s/SVM
MFCC	46	36	28
Delta-MFCC	37	33	26
CLS, $m = 1$	46	36	28
CLS, $m = 2$	34	23	18

Рис. 8: Ошибки в процентах на базе GTZAN, при использовании пятикратной кросс-валидации и для различного размера окон и классификаторов SVM и PCA.

Как и ожидалось процент ошибок для MFCC и первого порядка CLS совпадают, так как они измеряют похожие данные. Коэффициенты второго порядка CLS дают более высокую точность, так как они восстанавливают потерянную нестационарную информацию сигнала. Метод Delta-MFCC дает результаты лучше, чем MFCC, но он проигрывает CLS.

При увеличении T процент ошибок уменьшается. Однако на больших временных масштабах классификация страдает, поскольку даже CLS второго порядка не способен точно представить сигнал после реконструкции. Включение CLS третьего порядка (m = 3) незначительно улучшает результаты классификации, значительно увеличивая вычислительную нагрузку.

Замечание 3. *Последние результаты на базе GTZAN получены с классификаторами, лучше чем SVM, например, AdaBoost классификатор или Sparse Coding классификатор. Эти классификаторы могут быть также применены к CLS векторам для улучшения классификационных результатов.*

Литература

[1] S. Davis and P. Mermelstein Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, 1980.

[2] J. Anden, S.Mallat Multiscale Scattering for Audio Classification.Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011.

[3] S. Mallat Group Invariant Scattering, Communications in Pure and Applied Mathematics, Vol. 65 No. 10, 2012.

[4] J. Brown Computer identification of musical instruments using pattern recognition with cepstral coefficients as features, Journal of the Acoustical Society of America, Vol. 105, No. 3, pp. 1933–1941, 1999.

[5] Г. Фант Акустическая теория речеобразования. М.: Наука, 1964.