

КРИТЕРИИ ЭНТРОПИИ И ДИНАМИЗМА В ЗАДАЧАХ КЛАССИФИКАЦИИ РЕЧЕВЫХ И АУДИО ДОКУМЕНТОВ

Введение. В настоящее время очень интенсивно ведутся работы по созданию высокоинтенсивных систем автоматического распознавания слитной речи [1]. В ходе проведенных исследований было отмечено, что дальнейший прогресс в данном направлении невозможен без разработки новых инструментов, таких как анализ содержимого или индексация аудиоданных.

Для аудио- и мультимедиа информации вообще методы, основанные на точном соответствии, практически бесполезны. Следовательно, необходимо найти некоторую меру аудиоподобия. В литературе [2, 3] было предложено много векторов признаков для приложений классификации аудио. В этой статье описывается система классификации речевых и аудиодокументов, основанная на использовании таких характеристик, как энтропия и динамизм, которые впервые были представлены в работе [4].

Основная идея предложенного метода заключается в том, что входная нейронная сеть рассматривается в качестве информационного канала Канал, настроенный на определенный тип информации, пропускает ее лучше всего. В нашем случае в качестве такого информационного канала используется многослойный персептрон, выдающий апостериорные вероятности для распознавания речи. С помощью этих апостериорных вероятностей вычисляются два параметра: энтропия и динамизм. В качестве классификатора используется скрытая Марковская модель.

[4]

нейронной сети были использованы аллофоны русской речи. Благодаря этому была получена возможность не только отличать речь и музыку, но также и отделять русскую речь от других языков. Таким образом, метод хорошо работает также для идентификации языка.

Различные эксперименты показали эффективность возможностей использования критериев энтропии и динамизма не только для задач сегментации речь/музыка, но и для других приложений аудиоклассификации.

последовательность базовых дискретных сегментов. В качестве этих

сегментов могут быть использованы аллофоны, фонемы, дифоны. Предполагается, что эти фонологические элементы (например, фонемы) имеют особые артикуляционные и акустические характеристики. Фонемы могут быть описаны набором постоянных характеристик, называемых отличительными признаками. Эти признаки имеют прямое отношение к артикуляционному движению, при котором создаются речевые звуки, и отличаются твердо установленными акустическими параметрами. При разработке систем индексации речевых сигналов возникает проблема, как выбрать набор соответствующих параметров. Одно из лучших решений данной проблемы рассматривается ниже.

Предположим, что речевое высказывание в моменты времени $n = 1, 2, \dots, N$ представляется последовательностью наблюдаемых акустических векторов

$$X = \{x_1, x_2, \dots, x_N\}$$

где $x = \{v_b, \dots, v_d, \dots, v_D\}$ - акустический вектор, состоящий из D кепстральных коэффициентов.

Эта последовательность акустических векторов ассоциируется с последовательностью состояний

$$Q = \{q_1, q_2, \dots, q_N\}$$

где каждое состояние q_n принадлежит заданному множеству состояний.

Формально, для этих данных легко определить скрытую Марковскую модель в виде набора параметров

$$L = \{K, A, B, \pi\},$$

где K - число состояний модели; $A = \{a_{ij}\}$ матрица переходов, описывающая Марковскую цепь первого порядка. Здесь $a_{ij} = P(q_j/q_i)$ - вероятность перехода системы из состояния q_i в состояние q_j ; $B = \{b(x_n)\}$ - матрица вероятностей излучения, где $b(x_n) = P(x_n/q_i)$ - вероятность излучения вектора x_n в состоянии q_i ; π - матрица начальных состояний.

Как правило, для представления речевого высказывания используется некоторое множество фонем, которые представим в виде

$$\Phi = \{\phi_1, \dots, \phi_K, \dots, \phi_N\}$$

K -

пример, слово) представляется в виде последовательности фонем

$$\{\phi_1, \phi_2, \dots, \phi_N\}$$

Предположим, что с некоторым акустическим вектором $x \in X$ связана определенная вероятность $P(x/p)$, которая характеризует вероятность наблюдения вектора x , когда диктором произнесена фонема ($p \in \Phi$). Если обратиться к СММ, то можно отметить, что если в данной модели сопоставить каждому состоянию модели q_n определенную фонему (p_n), то между вероятностями $P(x/q_n)$ и $P(x/p_n)$ установится взаимосвязь. В СММ обычно для оценки локальной вероятности $P(x/q_n)$ используется три подхода:

1. Моделирование каждого класса фонем гауссовым распределением.
2. Моделирование каждого класса фонем смесью гауссовых

В работах [4, 5] было показано, что нейронная сеть может быть использована для оценки вероятности $P(p/x_n)$. Значит $P(p/x_n)$ указывает на то, что в момент времени n произнесена фонема (p) при условии, что наблюдается акустический вектор x_n .

Вероятности излучения, которые присущи СММ, могут быть получены при помощи теоремы Байеса:

$$P(x_n/p) = \frac{P(p/x_n)P(x_n)}{P(p)}$$

где вероятности $P(p/x_n)$ оцениваются при помощи нейронной сети, вероятности $P(p)$ оцениваются по базы данных, вероятности $P(x_n)$ полагаются постоянными.

Используя вероятности $P(p/x_n)$, которые оценивают при помощи именно:

энтропия

$$H(p) = - \sum_{i=1}^N \sum_{k=1}^K P(p_k/x) \log_2 P(p_k/x),$$

где N - число кадров в сегменте, n - номер кадра (индекс времени), $P(p_k/x)$ - апостериорная вероятность состояния (p_k в момент времени n).

Если сигнал похож на речь, то на выходе нейронной сети в каждый момент времени одна из вероятностей будет выше, и поэтому суммарная энтропия будет более низкой. И наоборот, энтропия будет выше для неречевых сигналов, где никакие фонемы не могут быть распознаны нейронной сетью.

- динамизм

$$= \sum_{j=1}^J \sum_{v=1}^{n+N/2} \sum_{m=1}^M \sum_{k=1}^K [P(p_{jk}/x) - P(p_{k|j-1})]^2$$

Этот параметр позволяет сравнивать апостериорные вероятности стериорные вероятности в случае речи изменяются намного быстрее, так

Структура предложенной системы. Полная блок-схема предложен-
. 1.

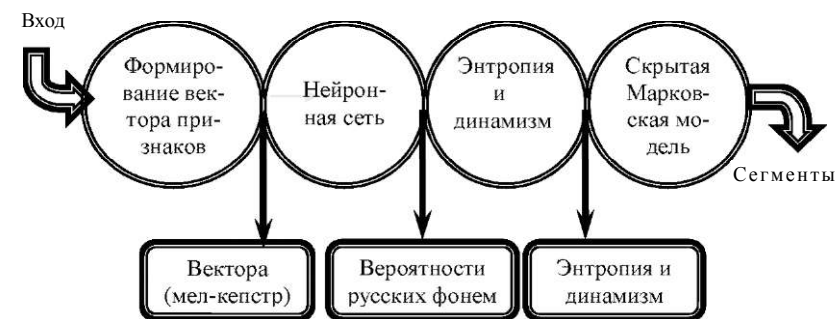


Рис. 1. Блок-схема предложенной системы

Формирование вектора признаков. Эта обработка включает следующие [5]:

- Предварительная обработка сигнала.
- Спектральный анализ.
- Кепстральный анализ

При проведении эксперимента используется вектор признаков, содержащий 12 кепстральных коэффициентов.

Структура нейронной сети. Персептрон - специальная разновидность нейронной сети, которая была выбрана для решения задачи классификации. Персептрон имел один скрытый слой, состоящий из 256 ней-

«мягкого максимума» (softmax) для выходного слоя. Дополнительные нейроны не вносили изменений в результат, а выбор сигмоидальной функции активации, а не функции тангенса или «мягкого максимума»,

Программная реализация и создание базы данных. Для достижения наибольшей производительности и быстродействия, программное обеспечение было реализовано с использованием модели компонентных объектов Microsoft (COM - Component Object Model). Система была создана как приложение Win32 (COM Сервер) при помощи языка программирования высокого уровня C ++.

Для проведения экспериментов по классификации аудиодокументов нами была создана база данных. Чтобы охватить все главные задачи ау-

25 языках. Для того чтобы сделать дальнейшее использование более

гласно международным кодам языков ISO 639. Затем была проведена ручная сегментация. Сегментированные файлы были помещены в различные папки, чтобы обеспечить возможность реализации различных

Наличие базы данных обеспечило возможность проведения различных экспериментов, например таких, как: сегментация музыка/речь/му-

висимая от языка), идентификация языка и распознавание диктора.

Результаты эксперимента. Для вычисления апостериорных вероятностей мы используем многослойный персептрон (входной слой 12 нейронов, скрытый - 256, выходной - 64) с функцией активации «мягкого максимума» для входного слоя, обученный посредством алгоритма обратного распространения ошибки. Входные параметры - первые 12 кепстральные коэффициента для спектра данных, оцифрованных с частотой выборки 16, использовалось окно 30 мс, которое сдвигалось на 10 мс. То есть в каждый момент времени на вход нейронной сети подается девять последовательных кадров. Для вычисления параметров: число аллофонов русской речи $K = 64$, размер окна усреднения $N = 40$.

Гистограмма энтропии для русского диктора показана на рис. 2 а. Так

онный канал, гистограмма энтропия напоминает нормальное распределение. Представление двух параметров (энтропии и динамизма) на одной плоскости для русского диктора показано на рис. 2 б. На рис. 2 в показав

ских дикторов. Все они фактически не отличаются как по среднему, так и по дисперсии.

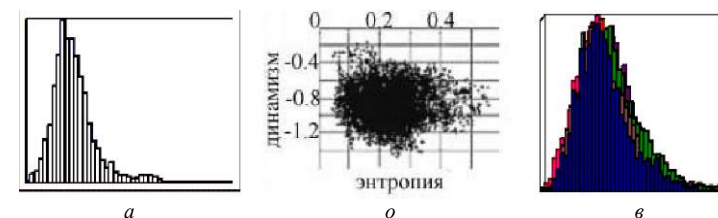


Рис. 2. Характеристики для русской речи: а - гистограмма энтропии для русского диктора; б - энтропия и динамизм для русского диктора; в - гистограммы энтропии различных русских дикторов

3

тограммы энтропии музыки различных стилей. Поскольку музыка не содержит аллофонов и не является тем типом сигнала, на который настроен наш канал, его гистограмма энтропии не похожа на нормальное распределение. Она отличается и по среднему и по дисперсии. Вместе энтропия и динамизм для русского диктора и музыки показаны на рис. 3 б. Таким образом, критерии энтропия и динамизм могут эффективно использоваться для задач сегментации речь/музыка.

Поскольку многослойный персептрон был обучен, используя аллофоны русской речи, то характерное поведение на его выходе охраняется только для русской речи. Набор аллофонов и их произношение различны для различных языков. И это сильно влияет на апостериорные вероятности и, следовательно, на энтропию и динамизм.

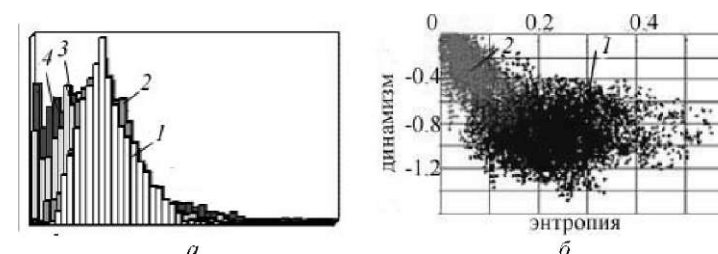


Рис. 3. Сравнение характеристик для речи и музыки: а - гистограммы энтропии для русских дикторов (1, 2) и гистограммы энтропии музыки различных стилей (3, 4); б - энтропия и динамизм для русского диктора (1) и музыки (2)

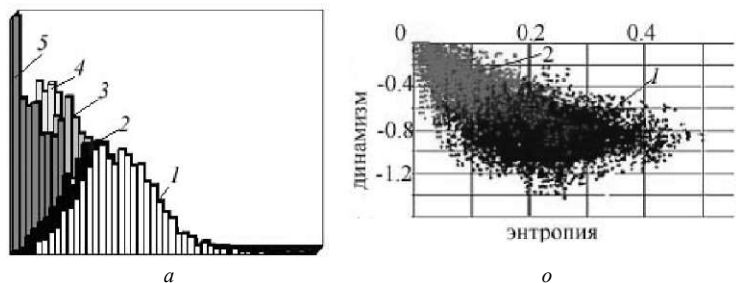
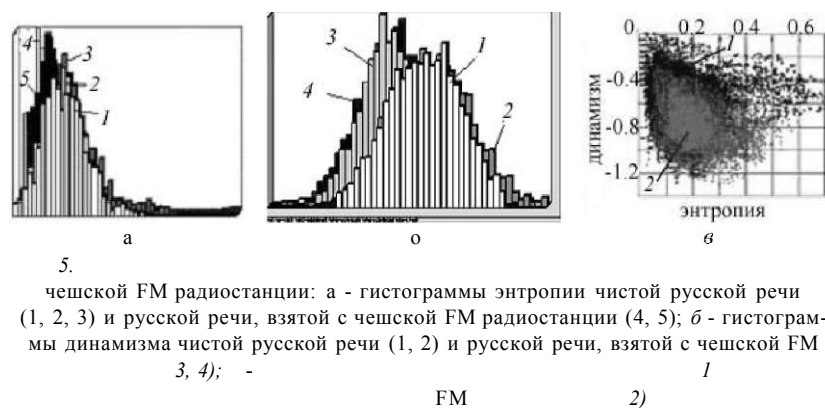


Рис. 4. Сравнение характеристик для русской и иностранной речи: а - гистограммы энтропии для русской (1, 2) и иностранной (3, 4, 5) речи; б - энтропия и динамизм для русского (1) и французского (2) дикторов



5. чешской FM радиостанции: а - гистограммы энтропии чистой русской речи (1, 2, 3) и русской речи, взятой с чешской FM радиостанции (4, 5); б - гистограммы динамизма чистой русской речи (1, 2) и русской речи, взятой с чешской FM (3, 4); -

FM

2)

Гистограммы энтропии для русской и иностранной речи (английский, французский и немецкий языки) показаны на рис. 4 а. Они отличаются и французского дикторов показаны на рис. 4 б.

Система может даже реагировать на акцент языка. Гистограммы энтропии родной русской речи и русской речи, взятой с чешской FM радиостанции, показаны на рис. 5 а. Гистограммы динамизма показаны на рис. 6 б. А вместе энтропия и динамизм показаны на рис. 5 в.

водительность системы. Эти параметры, основанные на апостериорных вероятностях, действительно являются хорошими дискриминантными характеристиками и подходят для высокоэффективной сегментации речь/музыка и для классификации речевого документа по языкам.

нантных характеристик были использованы параметры энтропия и динамизм, основанные на апостериорных вероятностях речевых фонетиче-

Система была протестирована на музыке различных стилей и речи на различных языках. Для проведения экспериментов была создана 350-часовая база данных радиопередач на 25 языках.

Наши результаты показывают, что вместе энтропия и динамизм, рассчитанные при помощи апостериорных вероятностей фонетических классов русской речи, являются мощным набором признаков для дискриминации речь/музыка и идентификации языка.

В результате предложенная система классификации речи и аудио

диопотоков, включая дискриминацию речь/музыка и идентификацию языка.

ЛИТЕРАТУРА

1. Morgan N., and Bourland H. Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach // Signal Processing Magazine, 1995. May. P. 25-42.
2. Williams G., Ellis D. Speech/music discrimination based on posterior probability features // Proc. of Eurospeech. 1999. P. 687-690.
3. Berenzweig A., Ellis D. Locating Singing Voice Segments within Music Signals // Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio. 2001. P.364-368.
4. Ajmera J., McCowan I., and Bourland H. Robust HMM-Based Speech/Music Segmentation // IDIAP Research Report RR 01-33. Martigny. Switzerland. 2001.
5. Boybel E., Kheidorov I., Chaikov Y. Wavelet-based biomedical signal processing using hidden Markov models // Proc. of 4th BSI International Workshop. 2002. Italy. P. 15-18.

лучшая селективная характеристика, чем динамизм. Как и ожидалось.