И. Э. ХЕЙДОРОВ, А. М. ЛУКАШЕВИЧ, Д. Л. МИТРОФАНОВ

КРИТЕРИИ ЭНТРОПИИ И ДИНАМИЗМА В ЗАДАЧАХ КЛАССИФИКАЦИИ РЕЧЕВЫХ И АУДИОДОКУМЕНТОВ

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.

В отличие от метода, описанного в работе [4], для обучения входной 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.

Выбор критерия. В основе произношения и восприятия речи лежит 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 $n = 1, 2, ..., N$ представляется последовательностью наблюдаемых акустических векторов

$$X = \{x_1, \ldots, x_n, \ldots, x_N\},$$

где $x = \{v_1, \ldots, v_d, \ldots, v_D\}$ – акустический вектор, состоящий из D кепстральных коэффициентов.

Эта последовательность акустических векторов ассоциируется с последовательностью состояний

$$Q = \{q_1, \ldots, q_n, \ldots q_N\},$$

где каждое состояние q_n принадлежит заданному множеству состояний.

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

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

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; $B = \{b_i(x_n)\}$ – матрица вероятностей излучения, где $b_j(x_n) = P(x_n/q_j)$ – вероятность излучения вектора x_n в состоянии q_j; π – матрица начальных состояний.

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

$$\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_\kappa, \ldots, \varphi_K\},$$

где K – число фонем.

Используя фонемы из данного множества, речевое высказывание (например, слово) представляется в виде последовательности фонем

$$\{\varphi_1, \varphi_2, \ldots, \varphi_n, \ldots, \varphi_N\}.$$

Предположим, что с некоторым акустическим вектором $x \in X$ связана определенная вероятность $P(x/\varphi)$, которая характеризует вероятность наблюдения вектора x, когда диктором произнесена фонема $\varphi \in \Phi$. Если обратиться к СММ, то можно отметить, что если в данной модели сопоставить каждому состоянию модели q_n определенную фонему φ_n, то между вероятностями $P(x/q_n)$ и $P(x/\varphi_n)$ 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 распределений.
3. Использование нейронной сети для оценки локальной вероятности.

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

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

$$P(x_n / \varphi_i) = \frac{P(\varphi_i / x_n) P(x_n)}{P(\varphi_i)},$$

где вероятности $P(\varphi_i/x_n)$ оцениваются при помощи нейронной сети, вероятности $P(\varphi_i)$ оцениваются из базы данных, вероятности $P(x_n)$ полагаются постоянными.

Используя вероятности $P(\varphi_i/x_n)$, которые оценивают при помощи многослойной нейронной сети, можно ввести две важные характеристики, а именно:

- энтропия

$$H_n = -\frac{1}{N} \sum_{t=n-N/2}^{n+N/2} \sum_{k=1}^{K} P(\varphi_k \mid x_t) \log_2 P(\varphi_k \mid x_t),$$

где N – число кадров в сегменте, n – номер кадра (индекс времени), $P(\varphi_k/x_t)$ – апостериорная вероятность состояния φ_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

$$D_n = \frac{1}{N} \sum_{t=n-N/2}^{n+N/2} \sum_{k=1}^{K} [P(\varphi_k \mid x_t) - P(\varphi_k \mid x_{t-1})]^2.$$

Этот параметр позволяет сравнивать апостериорные вероятности классов фонем в настоящий и предыдущий моменты времени. Эти апостериорные вероятности в случае речи изменяются намного быстрее, так что динамизм выше для речевых сигналов и ниже для неречевых.

Структура предложенной системы. Полная блок-схема предложенной системы аудиоклассификации показана на рис. 1.

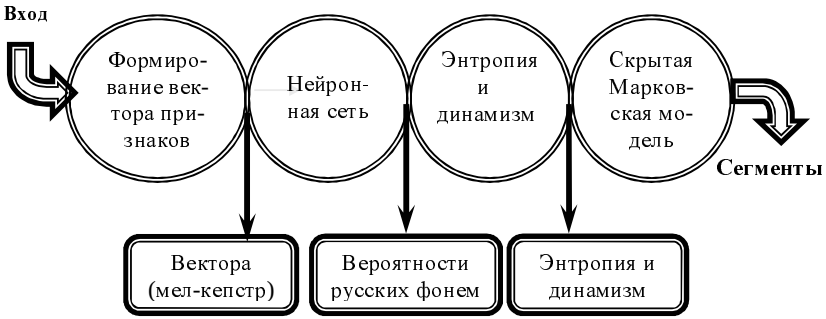


Рис. 1. Блок-схема предложенной системы

Формирование вектора признаков. Эта обработка включает следующие стадии [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 нейронов, сигмоидальную функцию активации для этого слоя и функцию «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», был обусловлен лучшей работой системы при данных параметрах.

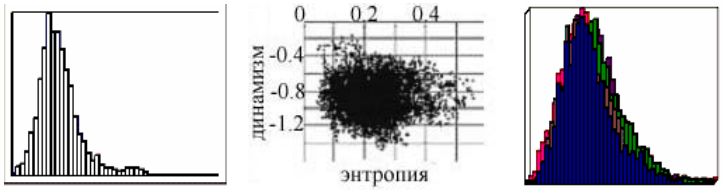
Программная реализация и создание базы данных. Для достижения наибольшей производительности и быстродействия, программное обеспечение было реализовано с использованием модели компонентных объектов Microsoft (COM – Component Object Model). Система была создана как приложение Win32 (COM Сервер) при помощи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

Для проведения экспериментов по классификации аудиодокументов нами была создана база данных. Чтобы охватить все главные задачи аудиоклассификации, было взято приблизительно 350 ч. радиопередач на 25 языках. Для того чтобы сделать дальнейшее использование более удобным, аудиофайлы были сохранены под различными именами согласно международным кодам языков ISO 639. Затем была проведена ручная сегментация. Сегментированные файлы были помещены в различные папки, чтобы обеспечить возможность реализации различных приложений классификации речи и аудио.

Наличие базы данных обеспечило возможность проведения различных экспериментов, например таких, как: сегментация музыка/речь/музыка с речью/тишина, сегментация мужчина/женщина (зависимая и независимая 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 как этот тип сигнала – точно тот, на который настроен наш информационный 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* показаны гистограммы энтропии различных, как мужских, так и женских русских дикторов. Все они фактически не отличаются как по среднему, так и по дисперсии.

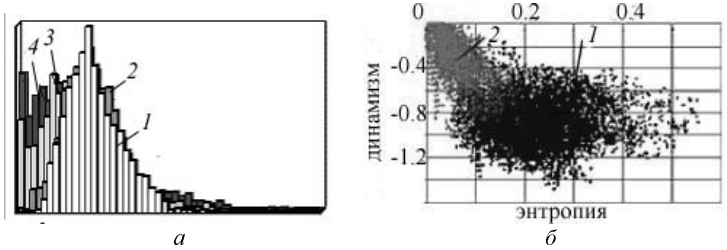


а *б* *в*

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

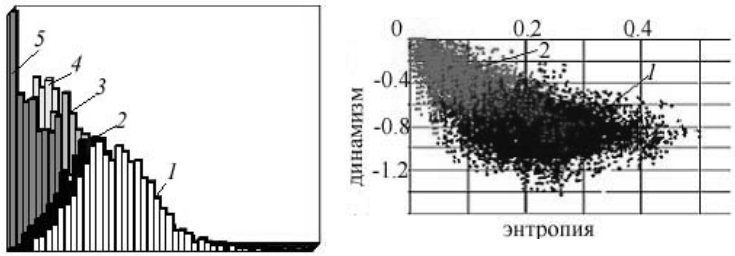
На рис.3*а* показана гистограмма энтропии для русского диктора и гистограммы энтропии музыки различных стилей. Поскольку музыка не содержит аллофонов и не является тем типом сигнала, на который настроен наш канал, его гистограмма энтропии не похожа на нормальное распределение. Она отличается и по среднему 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.



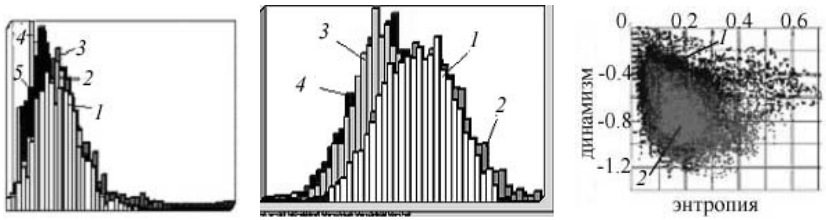
а *б*

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*)



а *б*

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



а *б* *в*

Рис. 5. Сравнение характеристик для чистой русской речи и русской речи, взятой с 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 радиостанции (*3, 4*); *в* – энтропия и динамизм чистой русской речи (*1*) и русской речи, взятой с чешской FM радиостанции (*2*)

Гистограммы энтропии для русской и иностранной речи (английский, французский и немецкий языки) показаны на рис. 4 *а*. Они отличаются и по среднему, и дисперсии. Вместе энтропия и динамизм для русского 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*.

Итак, энтропия и динамизм – хорошие критерии не только для дискриминации речи/музыки, но также и для идентификации языка.

Различные проведенные эксперименты показывают, что энтропия – лучшая селективная характеристика, чем динамизм. Как и ожидалось,

применение обоих параметров, энтропии и динамизма, улучшает производительность 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.

Заключение. В этой статье мы представили подход для решения задач классификации речевых и аудиодокументов. В качестве дискриминантных характеристик были использованы параметры энтропия и динамизм, основанные на апостериорных вероятностях речевых фонетических классов.

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 обеспечивает мощный, устойчивый метод для надежной сегментации аудиопотоков, 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., McCovan I., and Bourland H.* Robust HMM-Based Speech/Music Segmentation // IDIAP Research Report RR 01–33. Martigny. Switzerland. 2001.
5. *Bovbel 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.