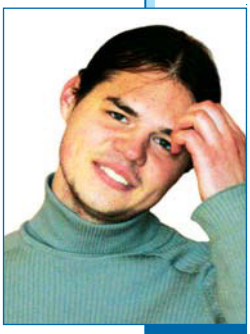
Классификация аудиосигналов с использованием одноклассового метода опорных векторов для систем поиска информации в мультимедиа-архивах

Янь Цзинбинь,
аспирант
У Ши,
аспирант
А.М. Сорока,
магистрант
А.А. Трус,
магистрант



В данной статье представлен способ классификации аудиосигналов на основе одноклассового метода опорных векторов (МОВ). Анализ полученных в ходе эксперимента результатов показывает, что данный метод имеет хорошую точность классификации, и его эффективность выше, чем у классификаторов на основе метода Байеса, скрытых Марковских моделей (СММ) и нейронной сети.

Abctract

This paper proposes an audio classification method based on one-class support vector machine (SVM). Experiment results show that SVM has good classification accuracy, and performs better than other classification systems using Bayes method, Hidden Markov Model (HMM) and Neural Network (NN).

Введение

Классификация аудиосигналов — это задача разделения непрерывного потока акустических данных на однородные участки (речь, музыка, звуки окружающей среды,

тишина и т.д.). С одной стороны, решение данной задачи можно использовать для удаления неречевых фрагментов из аудиосигналов, что приведет к увеличению скорости и точности распознавания речи. С другой стороны, сформулированная выше задача классификации является шагом предварительной обработки аудиосигналов в задачах классификации музыки по жанрам [1, 2]. Задача классификации является неотъемлемой частью алгоритмов индексации аудиосигналов и на сегодняшний день выполняется вручную, что приводит к огромным затратам человеческого труда и материальных ресурсов. Таким образом, построение системы автоматической классификации аудиосигналов является на сегодняшний день одним из наиболее актуальных нерешённых вопросов мультимедиа-технологий.

Традиционно алгоритм классификации аудиосигналов использует методы на основе решающего правила, учитывающие один или несколько признаков и принимающие решение путём сравнения с некоторым порогом, значение которого определяется обычно эмпирическим способом [3]. Такой подход имеет вполне очевидные недостатки. Во-первых, выбор решающего правила и последовательности классификации на группы не обязательно оптимальны. Во-вторых, ошибки верхнего уровня классификации переходят на следующий уровень и постепенно накапливаются, что значительно снижает общую точность системы. В-третьих, выбор порогового значения для решающего правила в значительной степени зависит от условий эксперимента; незначительное изменение качества речевых сигналов может привести к необходимости повторного обучения системы.

На протяжении последних лет ведутся поиски новых подходов и алгоритмов для классификации речевых сигналов: в частности, для этой задачи использовались классификаторы на основе метода K-ближайших соседей [4], нейронных сетей, моделей гауссовых смесей и алгоритма K-ближайших соседей. Однако каждому из этих методов в той или иной степени свойственны недостатки, не позволяющие строить на их основе высокоточные системы автоматической классификации аудиоданных.

В данной статье предлагается использовать новый подход к классификации аудиосигналов. Метод опорных векторов (МОВ) позволяет найти оптимальную гиперплоскость, разделяющую классы, в некотором модифицированном пространстве признаков, что позволяет преодолеть недостатки алгоритмов на основе решающих правил и пороговых значений. В качестве основной задачи рассматривается разделение аудиосигналов на пять основных классов: речь, речь с окружающими звуками, музыка, тишина и шум.

1. Построение вектора признаков аудиосигнала

При классификации аудиосигналов особое значение приобретает выбор способа построения для вектора признаков сигнала, которые обладают достаточной различающей способностью и являются устойчивыми.

Основные признаки, используемые для решения задачи классификации аудиосигналов, можно разделить на три большие группы. Первая группа признаков — это спектральные характеристики, отражающие свойства одного фрейма.

Признаки второй группы — комплексные спектральные характеристики, отражающие динамические свойства сигнала путём анализа нескольких соседних или всех фреймов сигнала в анализируемой области. К третьей группе признаков относятся характеристики сигнала, построенные на основе анализа последовательности временных отсчётов сигнала.

Для проведения классификации аудиосигналов важное значение имеют характеристики, в той или иной степени отражающие форму спектра. Частота основного тона (ЧОТ) является характерной и важной величиной для речевых сигналов и характеризует, в первую очередь, частоту колебаний голосовых связок человека. Формы ЧОТ для речи и музыки принципиально различаются в силу разной природы образования этих звуков. ЧОТ речи имеет сложную, пересечённую форму и включает лишь небольшой пропорциональный гармоничный состав. А для музыки характерна более гладкая форма ЧОТ.

Другой из таких характеристик является распределение энергии по спектральным диапазонам. Область частоты разделяется на четыре поддиапазона $\mathrm{sbi}(i=0,1,2,3)$, соответственно $[0, w_0/8],[w_0/8, w_0/4],[w_0/4, w_0/2],[w_0/2, w_0]$, и для каждого вычисляются значения SW_1, SW_2, SW_3, SW_4 согласно выражению (1):

$$SW_i = \frac{1}{E}\sum_{i=L_j}^{H_j}\left|f(i)\right|^2, \qquad (1)$$

где $f(i)$ — коэффициенты БПФ-преобразования данного фрейма, $w_0 = \frac{1}{2} f_s$ f — частота дискретизации, L_j и N_j — нижняя и верхняя частоты диапазона. Энергия аудиосигналов разных типов по-разному распределяется по диапазону. Музыка имеет более равномерное распределение энергии по диапазонам, речь практически целиком фокусируется в первом диапазоне (около 80%).

Помимо распределения энергии по диапазонам, для моделирования формы спектра фрейма важную роль играет центроид частоты, обозначаемый FC и определяемый согласно (2), а также ширина спектра (3).

$$FC = \frac{\sum_{i=0}^{w_0}\left|f(i)^2\right|\cdot i}{\sum_{i=0}^{w_0}\left|f(i)^2\right|} \qquad (2)$$

$$BW = \sqrt{\frac{\sum_{i-0}^{w_0}(i-FC)^2\left|f(i)^2\right|}{\sum_{i=0}^{w_0}\left|f(i)\right|^2}} \qquad (3)$$

В частности, для речевых сигналов, согласно (3), частоты находятся в диапазоне 0.1 кГц — 3.4 кГц, а для музыки характерен диапазон 0.02 кГц — 22.05 кГц.

Помимо характеристик, отражающих характеристики отдельных фреймов, для анализа аудиосигналов очень важно использовать комплексные характеристики, отвечающие

за поведение некоторых участков сигнала в целом. В частности, в качестве компоненты характеристического вектора можно использовать не точное значение кратковременной энергии, а некоторое модифицированное значение отклонения кратковременной энергии (МОКЭ), обозначаемое как $LSTER$ и определяемое следующей формулой:

$$\mathrm{LSTER} = \frac{1}{2N}\sum_{n=0}^{N-1}\left[\mathrm{sgn}(0.5avSTE - STE(n)) + 1\right] \quad , \qquad (4)$$

где $STE = \frac{1}{L}\sum_{i=0}^{L-1} x^2(i)$ — кратковременная энергия фрейма, L — длина анализируемого фрейма, $avSTE = \frac{1}{N}\sum_{n=0}^{N-1} STE(n)$ — усреднённая кратковременная энергия, N — общее число фреймов. МОКЭ является эффективной характеристикой для различения речевых и музыкальных сигналов. В общем случае, поскольку для речи характерно большое число фреймов, содержащих тишину, МОКЭ для речевых сигналов выше, чем для музыки, как показано на ***рис. 1***. На ***рис. 2*** представлено распределение вероятностей для МОКЭ речевого и музыкального сигнала. Если в качестве вектора признаков для разделения музыкальных и речевых сигналов использовать только МОКЭ, а в качестве порога принятия решения использовать точку пересечения двух кривых (рис. 2), то ошибка классификации составит всего порядка 8%.

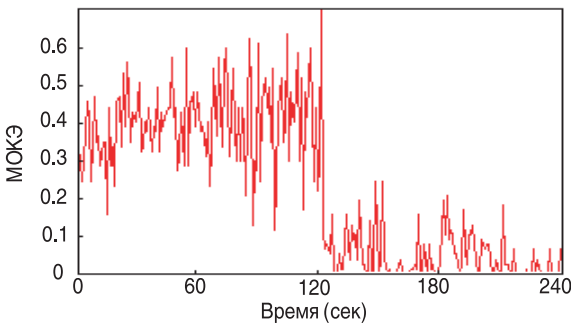


Рис. 1. Модифицированное значение отклонения кратковременной энергии (0–120 сек: речь, 121–240 сек: музыка)

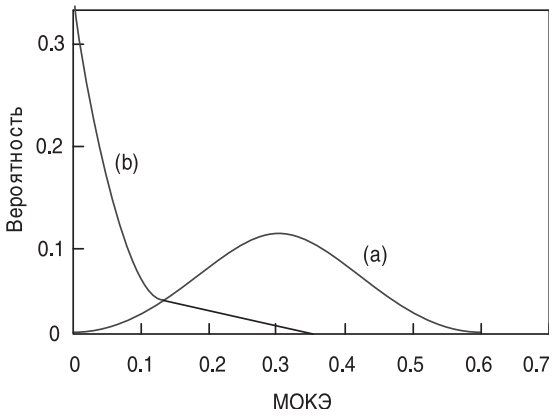


Рис. 2. Распределение вероятности МОКЭ: речь (а), музыка (б)

Другой комплексной характеристикой сигнала является спектральный поток, обозначаемый SF и определяемый как значение средней вариации спектра между двумя соседними фреймами в рамках одной секунды:

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}\left[\log(A(n,k)+\delta - \log(A(n-1,k)+\delta\right]^2 \quad ,$$

где $A(n, k)$ — дискретное преобразование Фурье (ДПФ) от n-го фрейма входного сигнала:

$$A(n,k) = \left| \sum_{m=-\infty}^{\infty} x(m) w(nL-m) e^{j\frac{2\pi}{L}km} \right| ,$$

где $x(m)$ — исходные входные данные, $w(m)$ — функция окна, L — длина окна, K — порядок ДПФ, N — общее число фреймов и S — очень малое значение, нужное, чтобы избежать переполнения разрядной сетки при вычислениях. В ходе экспериментов было установлено, что спектральный поток для речевых сигналов выше, чем для музыкальных ***(рис. 3)***. Речевой сегмент представлен с 0 по 120 секунду, музыка — с 121 по 240 секунду.

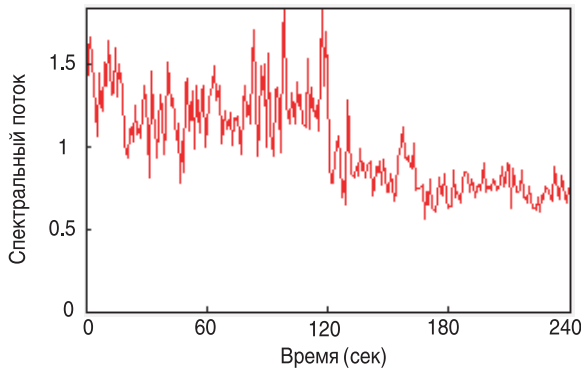


Рис. 3. Спектральный поток для речи (0–120 сек) и музыки (120–240 сек)

О временной структуре сигнала позволяет судить частота переходов через ноль (ЧПН), обозначаемая как ZCR и определяемая следующей формулой:

$$ZCR = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} \left| \mathrm{sgn}[x(m+1)] - \mathrm{sgn}[x(m)] \right| ,$$

где $x(m)$ — дискретный сигнал.

Соотношение тишины определяется как отношение количества фреймов с тишиной к суммарному количеству фреймов в сегменте и обозначается SR:

$$SR = \frac{количество \quad фреймов \quad тишины}{общее \quad число \quad фреймов}$$

В речи часто встречаются паузы, поэтому соответствующее соотношение тишины будет больше для речи по сравнению с музыкой.

ЧПН является важным параметром для описания различных аудиосигналов, однако в некоторых случаях более устойчивым параметром является модификация этой величины (МЧПН), определяемая следующим образом:

$$HZCR = \frac{1}{2N} \sum_{n=0}^{N-1} \left[\mathrm{sgn}\left(ZCR(n) - 1.5 \frac{1}{N} \sum_{n=0}^{N-1} ZCR(n) \right) + 1 \right] ,$$

где n — индекс фрагмента, N — общее количество фрагментов в окне длительностью в 1 секунду. В общем случае речевые сигналы содержат чередующиеся вокальные и невокальные звуки, в то время как такая структура не характерна для музыкальных сигналов. Таким образом, МЧПН будет выше для речевых сигналов по сравнению с музыкальными ***(рис. 4)***.

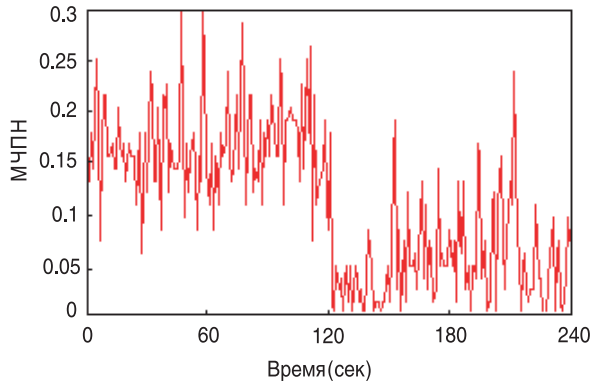


Рис. 4. Модифицированная частота переходов через ноль для речи (0–120 сек) и музыки (121–240 сек)

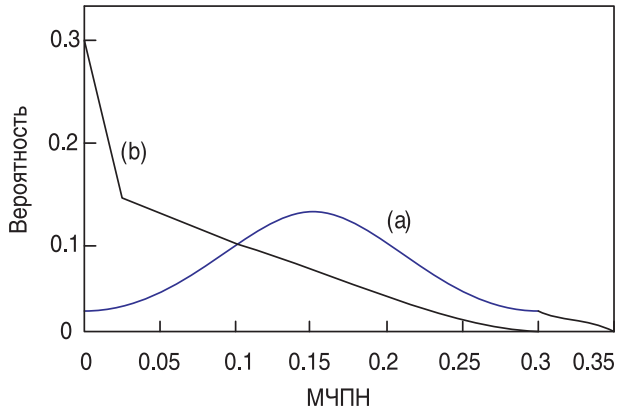


Рис. 5. Распределение вероятностей МЧПН для речи (a) и музыки (b) ноль для речи (0–120 сек) и музыки (121–240 сек)

На ***рис. 5*** представлены кривые распределения вероятностей МЧПН для речи и музыки. При использовании в качестве вектора признаков только МЧПН ошибка сегментации составит порядка 20%.

2. Частота основного тона (ЧОТ)

Традиционный способ определения частоты основного тона с использованием кепстра имеет ряд недостатков: высокая вычислительная сложность, значительное снижение точности определения при наличии фонового шума. Исследования показывают, что алгоритмы извлечения ЧОТ на основе вейвлет-анализа менее чувствительны к наличию фоновых шумов в сравнении с традиционными методами [6]. В качестве базовой функции непрерывного вейвлет-преобразования наиболее часто используется функция Morlet. Аналитическое выражение базовой функции непрерывного вейвлет-преобразования описывается следующей формулой:

$$\psi(x) = Ce^{-x^2/2}\cos(5x)$$

Поскольку большая часть энергии рассматриваемых сигналов лежит в низкочастотном диапазоне, используется логарифмическая шкала с целью улучшить частотное разрешение преобразования. Масштаб вейвлет-функций выбирается в пределах от 10 до 1024 и рассчитывается следующим образом:

$$scale_{k\in[1,num]} = MaxWvLng \cdot \exp\left[-\frac{k}{num}\cdot\log\left(\frac{MaxWvLng}{MinWvLng}\right)\right],$$

где $scale_k$ — масштаб вейвлета с индексом k, $MaxWvLng$ — наибольший масштаб вейвлета, MinWvLng — наименьший масштаб вейвлета, num — число масштабов.

В данном случае каждую компоненту вейвлет-вектора необходимо вычислять, используя массив с вейвлетом соответствующего масштаба. Компоненты вычисляются по следующей формуле:

$$CWT(pos,k) = \sqrt{rscale[k]}\cdot\sum_{i=0}^{MaxWvLng}\left(wave\left[pos-\frac{MaxWvLng}{2}+i\right]\cdot wvlt[k][i]\right), \quad (5)$$

где pos — текущая позиция (во времени), k — номер вейвлета, $rscale[k]$ — число, обратно пропорциональное масштабу вейвлета, $wave[]$ — массив со значениями выборок анализируемого сигнала, $wvlt[k][i]$ — i-ая выборка k-ого вейвлета, масштаб которого обратно пропорционален значению $rscale[k]$.

По формуле (5) рассчитывается двумерный массив вейвлет-коэффициентов в координатах времени-частоты (рис. 6). Максимальное значение вейвлет-коэффициентов является ЧОТ.

Как показано на ***рис. 6***, ЧОТ музыки изменяется более плавно в сравнении с ЧОТ речи. Дополнительно в качестве компонентов векторов признаков используются: дисперсия ЧОТ; степень гармоничности, определяемая как отношение количества фреймов в сегменте, ЧОТ которых не равна нулю, к общему их числу; процентное соотношение гладких сегментов к их общему количеству.

Описанные выше признаки включены в состав вектора признаков аудиосигнала и совместно с 12 мел-частотными кепстральными коэффициентами составили 25-компонентный вектор, который и использовался для классификации.

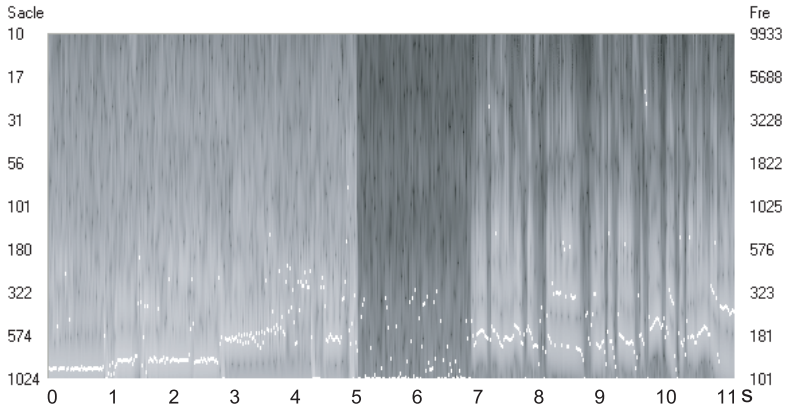


Рис. 6. Вейвлет-преобразование аудиосигнала (0-5 сек: музыка, 5.1-7 сек: тишина, 7.1-11 сек: речь)

3. Использование одноклассового метода опорных векторов в задачах многоклассовой классификации

Классический МОВ представляет собой бинарный классификатор. Для решения задач многоклассовой классификации конструируются каскады бинарных классификаторов. Это, в свою очередь, приводит к увеличению вычислительной сложности алгоритма, увеличению сложности настройки каждого классификатора, уменьшению точности итоговой модели классификации. Использование одноклассового МОВ позволяет увеличить точность и устойчивость итоговой модели классификации, а также снизить вычислительную сложность алгоритма обучения модели.

3.1. Одноклассовый метод опорных векторов

Предположим, что у нас имеется обучающая выборка $(x_1, y_1),...,(x_l, y_l)$, $x \in R^n$, $y \in \{+1,-1\}$, l — количество прецедентов в выборке, n — размерность пространства. В одноклассовом

методе опорных векторов задача обучения представляет собой задачу поиска гиперсферы минимального радиуса, включающей минимальное количество прецедентов из обучающей выборки. Задача обучения сводится к следующей оптимизационной задаче:

$$\min(\frac{1}{2}R^2+\frac{1}{vl}\sum_{i=1}^{l}\xi_i),,$$

где R — радиус гиперсферы, $\xi_1 > 0$ — штраф для соответствующего прецедента.

Набор параметров $\nu\,(0<\nu\le 1)$ представляет собой набор штрафов для каждого прецедента, и, таким образом, в гиперсферу может попадать некоторая часть прецедентов. При небольших значениях ν все прецеденты лежат на поверхности гиперсферы. Увеличение ν приводит к уменьшению радиуса гиперсферы. Решение задачи оптимизации может быть получено с использованием метода Лагранжа:

$$\min \quad W(\alpha)=\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K(x_i,x_j)\ ,$$

где $\sum_{i=1}^{l}\alpha_i=1$, $0\le\alpha_i\le\frac{1}{\nu m}$, $i=1,...,l$, $K(x_i,x_j)$ — ядерная функция.

Наконец, итоговая функция решения может быть представлена так:

$$f(x)=sign(\sum_{i=1}^{l}\alpha_i K(x_i,x)-\rho)\ .$$

Для любых $0<\alpha_i<\frac{1}{vm}$ и x_i рассчитывается параметр $\rho=\sum_{j=1}^{l}\alpha_j K(x_i,x_j)$.

3.2. Создание многоклассового классификатора на основе одноклассового метода опорных векторов

В задачах многоклассовой классификации может быть использован следующий подход: в процессе обучения для каждого класса строится решающее правило с использованием одноклассового МОВ. В процессе классификации проводится испытание тестовой выборки на полученных в процессе обучения функциях решения и в зависимости от максимального значения функции решения делается вывод о принадлежности выборки к какому-либо классу.

4. Экспериментальные результаты и анализ

4.1. База данных эксперимента

База данных эксперимента была представлена следующими аудиосигналами: чистая речь, музыка, звуки окружающей среды, речь с фоновым шумом и тишина.

Выборка речевых сигналов представлена аудиозаписями речи на четырёх различных языках. Выборка речевых сигналов с фоновыми шумами представлена аудиозаписями речи на фоне музыки, речи и городских звуков. Музыкальная выборка включает классическую детскую музыку, музыку, поп, джаз и другие виды.

Все данные представлены в виде аудиозаписей в формате PCM WAV 16kHz, 16Bit. Общий объём аудиоданных составил 500 минут, число сегментов равно 30000. Из каждого класса в качестве обучающей выборки использовалась 1/3 записей, в качестве тестовой выборки использовались 2/3 записей.

Точность классификации в ходе эксперимента определялась следующим образом:

$$точность = \frac{количество\ верно\ классифицированных\ выборок}{общее\ количество\ выборок}$$

4.2. Результаты

Результаты тестирования различных алгоритмов классификации представлены в таблице 1. Анализ полученных результатов показывает, что предложенный в данной статье подход позволяет получить более высокую точность классификации в сравнении с другими алгоритмами классификации.

Таблица 1

Результаты эксперимента

Тип аудиосигнала	Классификаторы			
	Метод Байеса	Скрытая Марковская модель	Нейронная сеть	Одноклассовый метод МОВ
Чистая речь	90.3%	94.7%	98.2%	98.7%
Музыка	82.5%	86.0%	91.7%	93.6%
Звуки окружающей среды	65.6%	68.1%	71.6%	75.5%
Речь с фоновым шумом	73.1%	78.3%	80.4%	85.0%
Тишина	91.1%	93.2%	94.5%	96.7%

5. Заключение

В данной статье рассмотрен подход к решению задачи классификации аудиосигналов с использованием одноклассового метода опорных векторов. Рассмотренный подход применён для классификации аудиосигналов, разделённых на пять классов: чистая речь, музыка, звуки окружающей среды, речь с фоновым шумом и тишина.

Анализ полученных результатов показывает, что использование одноклассового МОВ позволяет решить ряд проблем, присущих традиционным способам классификации, и получить модель классификации с более высокой точностью в сравнении с традиционными подходами.

Литература

1. *Foote J.* //Content-base retrieval of music and audio.In: Kuo C C J, et al(eds). Multimedia Storage and Archiving Systems II. Proc of SPIE, volume 3229, 1997. 138~147.
2. *Foote J.* //An overview of audio information retrieval. ACM-Springer Multimedia Systems, 1998.
3. *Srinivasan S., Petkovic D., Ponceleon D.* // Towards robust features for classifying audio in the cude Video system. In:Proc. of the 7 th ACM Intl. Conf. on Multimedia, Orlando, 1999. 393~400.
4. *Wold E., Blum T., Keislar D., Wheaton J.* // Content-based classification, search and retrieval of audio. IEEE Multimedia Magazine, 1996. 3(3): 27~36.
5. *Liu Z., Huang J., Wang Y., Chen T.* // Audio feature extraction and analysis for scene classification. In:IEEE Single Processing Society 1997 Workshop on Multimedia signal Processing.
6. *Dong Jing, Zhao Xiaohui, Ying Na.* Pitch detection algorithm based on dyadic wavelet transforms, Journal of Jilin University, 2006. 36(6)978-981.

Янь Цзинбинь —

аспирант БГУ, научные интересы — методы и алгоритмы обработки речевых сигналов, теория метода опорных векторов, алгоритмы обнаружения ключевых слов в потоке речи.

У Ши —

аспирант БГУ, научные интересы- методы и алгоритмы обработки речевых сигналов, теория метода опорных векторов, распознавание болезней голосового тракта по голосу

Сорока Александр Михайлович —

магистрант БГУ, научные интересы- методы и алгоритмы обработки цифровых сигналов, теория метода опорных векторов, смешанные гауссовы модели

Трус Александр Александрович —

магистрант БГУ, научные интересы- методы и алгоритмы обработки речевых сигналов, скрытые марковские модели, теория метода опорных векторов