ВЫБОР ОПТИМАЛЬНОГО НАБОРА ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ КЛАССИФИКАЦИИ ЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ДИКТОРА ПО ГОЛОСУ

Давыдов А. Г. (davydov-a@speetech.by),
Киселёв В. В. (kiselev-v@speetech.by),
Кочетков Д. С. (kochetkov-d@speetech.by),
Ткаченя А. В. (tkachenia-a@speetech.by)

ООО «Речевые технологии», Минск, Беларусь

Рассматривается вопрос минимизации количества информативных признаков при решении задачи классификации эмоционального состояния диктора по голосу. Доклад состоит из четырех частей. Первая часть содержит обзор информативных признаков характеризующих речь. Показывается, что акустические характеристики голоса могут быть условно разделены на пять категорий: просодические, динамические, фонационные, спектральные и энергетические. Во второй части кратко рассматриваются способы определения минимального эффективного набора информативных признаков. В экспериментальной части рассматривается процедура выбора оптимального набора информативных признаков на основе алгоритма sequential feature selection (SFS) с использованием мультиклассового SVM-классификатора для оценки эффективности. Приводятся результаты, полученные в ходе тестирования на берлинском корпусе эмоциональной речи. В четвертом разделе приводятся оценки эффективности классификации (порядка 85 % распознавания на берлинском корпусе для набора из 20 информативных признаков), кратко перечисляются наиболее эффективные информативные признаки и делаются предположения о возможных способах дальнейшего улучшения созданного классификатора.

Ключевые слова: классификация эмоций по голосу, информативные признаки, выбор информативных признаков

OPTIMAL FEATURE SELECTION FOR SPEAKER'S EMOTIONAL STATE CLASSIFICATION

Davydov A. G. (davydov-a@speetech.by),
Kiselev V. V. (kiselev-v@speetech.by),
Kochetkov D. S. (kochetkov-d@speetech.by),
Tkachenia A. V. (tkachenia-a@speetech.by)

LLC «Speech technologies», Minsk, Belarus

The research focus of present work is optimization of a feature set for voice emotion recognition. First part of the article contains a brief review of the most common speech features widely used in the emotion recognition tasks. It is shown that acoustic characteristics of a voice can be divided into five categories: prosodic, dynamic, qualitative, spectral and power. Also a number of the most effective features and statistical functionals derived from these features are discussed. After that two most widespread techniques of robust feature selection are explained. In the experimental part of the article, the feature selection algorithm developed on the basis of sequential feature selection (SFS) approach is presented. Further cross-validation procedure used in our studies is described. The recognition rate obtained on the Berlin database of emotional speech using optimized feature set (20 features) from the 10-fold cross-validation procedure is approximately 85%. In the conclusion we discuss some properties of the derived feature set and confusion matrix of the developed SVM classifier.

Key words: emotion state classification, feature vectors, feature selection

1. Обзор информативных признаков

Важнейшим звеном системы автоматического детектирования эмоций по голосу диктора является выделение оптимального набора информативных признаков, коррелированных с эмоциональными состояниями. Выбор информативных признаков оказывает значительное влияние на эффективность классификации. Условно характеристики речи диктора можно разбить на два основных класса — акустические и лингвистические. В зависимости от решаемой задачи их относительная эффективность может быть различной. В модельных эмоциональных базах, в которых речь дикторов соответствует заранее определенному сценарию, либо в многоязычных корпусах, на первый план выходят акустические параметры, в то время как при работе со спонтанной речью роль лингвистических признаков может оказаться весьма существенной [Schuller *et al.*, 2011].

Из-за нестационарности речевых сигналов во времени для определения их акустических характеристик записи обычно приходится разбивать

на небольшие фрагменты, именуемые фреймами. Предполагается, что в их пределах исходный речевой сигнал является квазистационарным. Характеристики речевого сигнала, определенные для каждого фрейма, называются локальными. Однако возможна работа и на уровне интегральных характеристик сигнала. Они определяются путем приложения некоторых статистических функционалов ко всем параметрам, выделенным из сигнала.

Однозначного ответа на вопрос, что лучше подходит для нужд распознавания — локальные или интегральные характеристики, пока нет. Предполагается, что применение интегральных параметров позволяет добиться более высокой точности и скорости классификации. Интегральные характеристики речевого сигнала наиболее эффективны для распознавания состояний, соответствующих различным уровням активации психики. Если же эмоциональные состояния соответствуют примерно одному уровню активности, различаясь, например, по валентности (гнев и радость), то их идентификация посредством анализа интегральных характеристик речевого сигнала окажется затруднена. Кроме этого, недостатками применения интегральных параметров является потеря информации об изменениях речевого сигнала во времени и невозможность применения сложных методов классификации из-за нехватки обучающих векторов. Использование локальных характеристик сигнала позволяет обойти большинство из вышеперечисленных трудностей.

Ранее при разработке систем детектирования эмоционального состояния диктора использовались в основном небольшие наборы информативных признаков (несколько десятков). Однако с течением времени, число выделяемых из звукового сигнала характеристик значительно возросло. Создаваемые на основе опыта экспертов признаки по-прежнему играют важную роль при разработке систем детектирования эмоций, однако все больше внимания уделяется методу «грубой силы» (*brute force extracting*). При этом число выделяемых информативных признаков достигает нескольких сотен, после чего к ним может быть дополнительно применен иерархичный набор функционалов. При таком подходе важнейшее значение приобретают последующие процедуры выбора оптимального набора характеристик, требуемых для эффективного обучения классификатора.

Акустические характеристики голоса могут быть условно разделены на пять категории [El Ayadi *et al.*, 2011]: просодические (частота основного тона, темп речи и т.д.), динамические (фонетическая функция [А. С. Рылов, 2003]), фонационные (отношение гармоник основного тона к шуму, джиттер, шиммер [M. Farrus *et al.*, 2008], …), спектральные (линейные спектральные частоты, кепстральные коэффициенты линейной шкалы частот, кепстральные коэффициенты мел-шкалы частот, …) и энергетические (отношение мощностей в спектральных полосах, оценка мощности сигнала и другие как правило, основанные на энергетическом операторе Тигера). Каждая группа показателей предназначена для описания отдельных аспектов голоса, и находит свое применение в распознавании эмоциональных состояний.

Выбор эффективных информативных признаков для разделения эмоциональных состояний диктора в значительной степени определяется перечнем детектируемых классов эмоциональной речи: характеристики, основанные

на энергетическом операторе Тигера, лучше всего приспособлены для детектирования состояния стресса; для разделения состояний с высоким и низким уровнем возбуждения лучше всего подходят просодические характеристики, такие как частота основного тона и энергия сигнала. Для различения нескольких эмоциональных классов необходимо задействовать спектральные параметры, которые также можно скомбинировать с просодическими характеристиками сигнала. Более точно наиболее эффективный набор информативных признаков может быть определен только на этапе экспериментального исследования.

Перед применением функционалов характеристики нижнего уровня можно подвергнуть фильтрации, трансформациям, можно вычислить их первые и вторые производные. Затем можно применять статистические функционалы, методы аппроксимации кривых, или основанные на специфике человеческого восприятия трансформации. Наиболее популярные статистические функционалы это четыре первых момента (среднее, стандартное отклонение, асимметрия и эксцесс), порядковая статистика (экстремумы и темпоральная информация о них), квартили, диапазоны амплитуд, скорость пересечения нуля, подъем/спад, начало/конец и анализ высшего порядка [Schuller *et al.*, 2011]. При помощи методов аппроксимации (как правило, линейной) кривых можно получить коэффициенты регрессии, например, коэффициент наклона и ошибки линейной регрессии.

На основе проведенного анализа существующих подходов классификации психоэмоционального состояния диктора по его речи были исследованы следующие базовые информативные признаки речи:

- просодические: частота основного тона, темп речи;
- динамические: фонетическая функция;
- фонационные: отношение гармоник основного тона к шуму, джиттер, шиммер;
- спектральные: линейные спектральные частоты, кепстральные коэффициенты линейной шкалы частот, кепстральные коэффициенты мел-шкалы частот;
- энергетические: отношение мощностей в спектральных полосах, оценка мощности сигнала.

С применением к ним следующих функционалов от базовых признаков: статистики высших порядков, первая и вторая производные, поведение кратковременных квантилей, энергетический оператор Тигера.

2. Выбор наиболее информативных и помехоустойчивых параметров речи

Число всевозможных информативных признаков, выделяемых из звукового сигнала, может достигать нескольких тысяч. Далеко не все из них эффективны для решения задач распознавания эмоционального состояния, а немалая часть потенциально полезных характеристик оказывается избыточными. Перед построением и обучением классификаторов необходимо провести предварительную процедуру отбора информативных признаков. Конечной целью

данного этапа работы является выделение релевантного набора характеристик звукового сигнала и декорреляция пространства информативных признаков.

В ранних работах информативные признаки выбирались эвристическими методами, с опорой на опыт экспертов. С течением времени появилась возможность задействовать для этих целей возросшую вычислительную мощность используемого аппаратного обеспечения.

Алгоритмы отбора информативных признаков, чаще всего используемые при разработке систем распознавания эмоционального состояния по голосу, можно условно разделить на два класса — «обертки» (*wrappers*) и «фильтры» (*filters*).

Отбор признаков с использованием «обертки» (*wrapper based selection*) использует оценку работы классификатора в качестве критерия оптимизации внутри замкнутого цикла. К сожалению, даже для сравнительно небольших объемов данных исчерпывающий перебор информативных признаков неприемлем. Необходимо выбрать более приемлемую с точки зрения вычислительной сложности, а значит и более ограниченную и менее оптимальную процедуру перебора. Вероятно, на данный момент наиболее популярной является стратегия линейного последовательного поиска (*sequential forward search)* [Pudil *et al.*, 1994]. Алгоритм начинает с пустого множества и последовательно добавляет к нему наилучшие информативные признаки.

Помимо «оберток», можно использовать «фильтры», к примеру, основанные на методах теории информации либо корреляционном анализе [Hall, 1998]. При этом критерием оптимизации является некоторая функция, соотносящаяся с корреляциями между информативными признаками, приростом информации при их добавлении к набору, определенными метриками в пространстве признаков, статистиками и т. п. Плюсом такого подхода является пониженная, по сравнению с «оберткой», вычислительная сложность алгоритма. В то же время, эффективность работы «фильтра», как правило, ниже, чем «обертки», так как предположения, положенные в его основу, будучи целесообразными в одних случаях, могут существенно нарушаться в других.

3. Эксперимент

3.1. Работа алгоритма автоматического выбора оптимального набора информативных признаков

В результате анализа литературы и предварительно проведенных экспериментов было принято решение использовать в процедуре выбора оптимального набора информативных признаков алгоритм sequential forward feature selection (SFFS). Суть его работы заключается в том, что на каждой итерации к набору добавляется признак, обеспечивающий наилучшую (для данной итерации) эффективность распознавания.

Мультиклассовый SVM-классификатор строится с использованием библиотеки libSVM[1], в которой данная опция реализована посредством набора классификаторов one-vs-one с последующим голосованием. Это позволяет сразу выбрать наиболее оптимальный набор информативных признаков именно для мультиклассовой классификации, не занимаясь подбором параметров для отдельных классов.

В SVM-классификаторе используется RBF-ядро, а в качестве информативных признаков используются расстояния (интеграл модуля разности маргинальных функций распределения) между функциями распределения различных видов наблюдений и соответствующих медианных функций распределения для различных эмоциональных классов.

После подбора очередного информативного признака параметры модели, используемой при обучении классификатора, корректируются для достижения оптимальной эффективности распознавания.

3.2. Оценка эффективности классификации

Эффективность распознавания на каждой итерации и в процедуре подбора параметров модели оценивается при помощи метода v-fold cross-validation, реализованного в составе пакета libSVM. Данный метод подразумевает разбиение обучающего множества на v частей, с последующим обучением классификаторов на v-1 части, тестированием на оставшейся части и усреднением полученных результатов.

Итоговая оценка сформированного набора информативных признаков определяется как медианное значение эффективности классификации, достигаемой моделью при случайном разделении исходного множества на обучающую и тестовую выборки.

3.3. Тестирование алгоритма SFFS на берлинском корпусе эмоциональной речи[2]

Тестирование алгоритма проводилось на записях, взятых из берлинского корпуса эмоциональной речи (Emo-DB). База включает 535 записей речи 10 дикторов (5 мужчин, 5 женщин), воспроизводящих набор дискретных эмоциональных состояний, называемых иногда «архетипическими» (гнев, раздражение, страх, радость, скука, печаль и нейтральное состояние). Авторское исследование

---

[1] libSVM — это программный пакет предназначенный для разработки алгоритмов на основе метода опорных векторов;

[2] Берлинский корпус эмоциональной речи — это речевая база, которая содержит записи эмоциональной речи на немецком языке, записанные профессиональными актерами в акустической комнате.

Берлинской базы показало [Burkhardt F *et al.*, 2005], что эмоции в ней распознаются слушателями в 80% случаев, и в 60% признаются естественными.

Зависимость эффективности распознавания эмоциональных состояний от числа добавленных в набор алгоритмом SFS информативных признаков показана на рисунке 1.

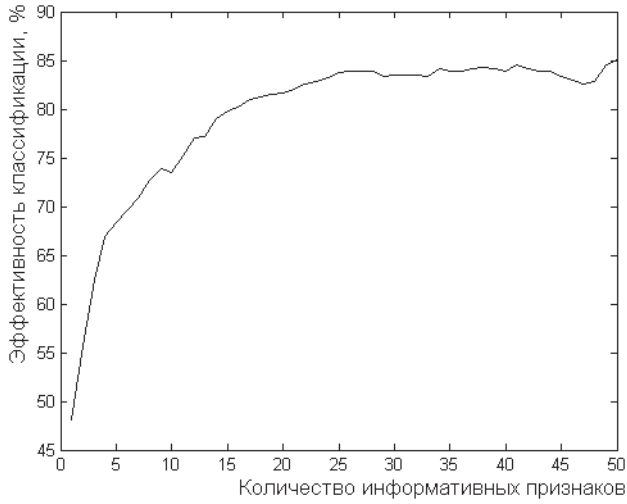

Рис. 1. Эффективность классификации как функция числа добавленных алгоритмом SFS информативных признаков

Видно, что на каждой итерации эффективность распознавания эмоциональных классов возрастала до тех пор, пока не достигла некоторого максимума. Затем эффективность распознавания возрастала с ростом количества признаков очень медленно (Рис. 1).

Тестирование, полученного в результате работы SFS, набора из 50 информативных признаков и подобранных параметров SVM показало, что эффективность классификации составляет приблизительно 85% (матрица спутывания приведена в таблице 1).

Таблица 1. Матрица спутывания классификации эмоциональных состояний

	гнев	скука	раздражение	страх	радость	нейтральное	печаль
гнев	0.8462	0	0	0.0769	0.0769	0	0
скука	0	0.8750	0	0	0	0.1250	0
раздражение	0	0	0.6000	0.4000	0	0	0
страх	0	0	0	1	0	0	0
радость	0.1429	0	0	0.1429	0.7143	0	0
нейтральное	0	0.1250	0	0	0	0.8750	0
печаль	0	0	0	0	0	0	1

4. Анализ и выводы

Выбор набора наиболее существенных информативных признаков при помощи sequential feature selection, а так же подбор оптимальных параметров модели SVM-классификатора позволил достичь эффективности распознавания эмоциональных состояний на корпусе Emo-DB порядка 85 %, что сравнимо с таковой для человека.

Для достижения указанной эффективности распознавания, согласно графику, приведенному на рисунке 1, достаточно 20 информативных признаков.

В число наиболее эффективных информативных признаков вошли: производная мощности, первая и вторая производные частоты основного тона, MFCC, LPCC и LSF коэффициенты, отношение мощностей в спектральных полосах, а также некоторые функционалы от указанных параметров.

Дальнейшее совершенствование классификатора представляется авторам в использовании динамических моделей эмоциональной речи.

References

1. *Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W. and Weiss B.* A Database of German Emotional Speech // Proc. Interspeech, 2005.
2. *El Ayadi, M., Kamel, M. S. and Karray, F.*, (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), pp. 572–587.
3. *Рылов, А. С.* Анализ речи в распознающих системах. — Мн.: Бестпринт, 2003. — 264 с.
4. *Farrus, M., Hernando, J.* (2008) Using Jitter and Shimmer in speaker verification. IET Signal Process., 2009, Vol. 3, Iss. 4, pp. 247–257.
5. *Hall, M. A.* (1998) Correlation-based feature selection for machine learning. Ph. D. Thesis, Hamilton, NZ: Waikato University, Department of Computer Science.
6. *Pudil, P., Novovicova, J., Kittler, J.* (1994) Floating search methods in feature selection. Pattern Recognition Lett. 15, pp. 1119–1125.
7. *Schuller, B., Batliner, A., Steidl, S. and Seppi, D.,* (2011) Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, In Press.