ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Фундаментальная и компьютерная лингвистика»

Сонина Полина Алексеевна

СРАВНЕНИЕ РАБОТЫ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ: СПЕКТРОГРАММА И ЗВУК

Выпускная квалификационная работа

студентки 4 курса бакалавриата группы БКЛ141

Академический руководитель
образовательной программы
канд. филологических наук, доц.
Ю.А. Ландер

_____

« » _____ 2018 г.

Научный руководитель

старший преподаватель
Г.А. Мороз

_____

« » _____ 2018 г.

Москва 2018

Оглавление

1. Введение

На данный момент разработка систем для автоматического анализа устной речи является одним из основных направлений в области машинного обучения. Системы распознавания речи широко применяются для извлечения текстовой информации из аудио как в научных целях, так и в коммерческих.

Чтобы аудио данные можно было использовать при обучении модели, нужно провести извлечение признаков (feature extraction) из первоначального сигнала. Для извлечения признаков из аудио существует набор стандартных метрик, которые используются во многих существующих системах. Высокое качество работы современных систем обычно достигается за счёт обучения на большом количестве данных, а также за счёт доступности больших вычислительных ресурсов.

Существуют исследования, предлагающие альтернативный подход к анализу аудио данных. Подход заключается в представлении аудио в виде спектрограммы - изображения, показывающего зависимость спектральной плотности мощности сигнала от времени. Использование спектрограмм для классификации речевых единиц представляется возможным благодаря тому, что форманты, образующие речевые единицы, различимы на спектрограмме для человеческого глаза. Подобные исследования в большинстве своём проведены на англоязычных данных, для русскоязычных же данных конкретный подход недостаточно исследован.

Цель данной работы – исследовать подход к извлечению признаков, основанный на методах классификации изображений, и сравнить его работу с общепринятыми для аудио подходами. Так же, целью являлось максимизировать точность работы алгоритма классификации на малом объёме данных при небольших вычислительных затратах.

Задачи исследования состояли в:

- сборе небольшого количества аудио данных на русском языке;
- создании алгоритма для классификации данных, использующих стандартные способы извлечения признаков;

- создании алгоритма, где аудио данные представляются в виде спектрограмм, к которым применяются методы извлечения признаков из изображений;
- сравнении работы полученных алгоритмов;
- анализ ошибок, характерных для каждой из систем.

2. Обзор литературы

Современные исследовательские работы по распознаванию речи существуют в огромном количестве. Однако статьи, посвященные использованию обработки изображений спектрограмм, особенно на русскоязычных данных встречаются намного реже.

Статья (Washani, Sharma 2015) дает обзор области исследований распознавания речи, существующие проблемы, прогресс и текущее состояние. В статье подчеркивается важность этапа предварительной обработки для борьбы с наиболее значимой проблемой - шумом.

(Madan, Gupta 2014) рассматривает методы обработки речи, фокусируясь при этом на методах извлечения признаков и классификации данных. Методы, сравниваемые в статье, являются наиболее используемыми в этой области и будут полезны для нашего исследования. Мел-кепстральные коэффициенты (MFCC) в сочетании с другими методами предлагаются для извлечения признаков, а алгоритм скрытых марковских моделей - для классификации слов в речевом потоке.

(Dennis 2014) предлагает подход к распознаванию звуковых событий, основанный на анализе изображений. Данная работа развивает метод, основанный на извлечении признаков из спектрограммы в сочетании с более традиционными методами обработки звука. Результаты оказались довольно успешными, но исследование было сосредоточено на классификации звуковых событий, а не на распознавании речи. Однако, разработка подобных моделей для распознавания речи предлагается автором как перспективная область разработки.

Статья (Nguyen, Bui 2016) предлагает модель, основанную на извлечении признаков из спектральных изображений для классификации речи. Конкретный подход также позволяет добавлять обучающие данные без переобучения модели,

что полезно для задач обработки больших объёмов данных. Модель показала удовлетворительные результаты на различных наборах данных.

3. Данные

Для обучения экспериментальных моделей были собраны аудио данные в формате wav. На записях носители русского языка произносили числа от одного до пятидесяти (1-50). В результате было использовано 27 наборов данных, состоящих из 50 речевых единиц. Говорящие на записях – взрослые (21-55 лет) жители Москвы и Московской области, мужчины и женщины. Для записи были использованы встроенные микрофоны нескольких смартфонов, поэтому на качество аудио повлияли различия в чувствительности устройств. Также, запись проводилась в разных условиях, что привело к различным уровням зашумлённости данных. Качество собранных данных близко бытовым условиям использования технологии распознавания речи, например, в качестве пользовательских интерфейсов.

Данные, а также коды программ расположены в репозитории GitHub (ссылка в Приложении).

4. Этапы работы

Программы для работы с данными были написаны на языке Python 3. Были использованы дополнительные библиотеки:

- numpy, os, time, matplotlib – для общих задач;
- librosa – для работы с аудио данными;
- scipy, scikit-image – для работы с изображениями;
- scikit-learn – для создания и оценки обучаемых моделей;
- tensorflow – для создания искусственных нейронных сетей.

4.1. Подготовка данных

Скрипт для создания наборов данных из записей поочерёдно загружает аудиозаписи из указанной папки и разбивает их на 50 частей, избавляясь от пауз. Для разбиения на сегменты используется функция split из библиотеки librosa. В функцию передаются параметры top_dp – граница в децибелах относительно

звукового сигнала, ниже которой фрагмент будет считаться тишиной; и frame_length – длина анализируемых фреймов. Параметры настраивались вручную для аудио, записанных в различных условиях. Полученные аудио фрагменты сохраняются, а в названии каждого файла указывается метка класса - произнесённое число.

4.2 Классификация аудио

Все последующие эксперименты выполнены в виде блокнотов ipynb, содержащих фрагменты кода, выдачи фрагментов и текстовые примечания.

Программа загружает размеченные сегменты аудио. Затем с помощью функций из библиотеки librosa из каждого сегмента извлекаются признаки:

- мел-кепстральные коэффициенты (MFCC);
- спектрограмма в формате мел-шкалы (mel-scaled power spectrogram);
- хромограмма кратковременного преобразования Фурье (chromagram of a short-time Fourier transform);
- спектральный контраст октавы (octave-based spectral contrast).

Средние значения каждой из полученных матриц собираются в вектор признаков (feature vector). Создаётся массив векторов и массив соответствующих им маркеров класса (1-50). Данные разделяются случайным образом на обучающую и тестовую выборку (распределение 25 к 2).

В первую очередь для классификации используются достаточно универсальные и быстрые алгоритмы: RandomForestClassifier (случайные деревья) и SVM (метод опорных векторов). Реализации данных методов есть в составе библиотеки scikit-learn. Для классификатора RandomForestClassifier был подобрано значение параметра n_estimators (число деревьев) – 500, при дальнейшем повышении значения параметра точность работы модели не возрастала. Для SVM подобрано значение параметра kernel (тип ядра) – poly (полиномиальное), при котором получен лучший результат.

При помощи библиотеки TensorFlow создана достаточно простая по структуре нейронная сеть с двумя скрытыми уровнями нейронов. На первом уровне используется функция активации tanh (гиперболическая), а на втором уровне sigmoid (сигмоида). Модель применялась в двух вариациях: по 300 нейронов на каждом уровне, и по 500 нейронов для повышения точности. Количество эпох

обучения для каждой модели составляло – 50000, а шаг обучения (learning rate) – 0,00001.

Также были созданы модели RandomForestClassifier и SVM, обученные только на векторе, созданном на основе самой широко используемой метрики – MFCC, но без усреднения получаемой матрицы.

4.3 Классификация спектрограмм

Следующим этапом было создание метода извлечения признаков из аудио, основанного на подходе к классификации изображений. Для каждого фрагмента аудио формируется мел-спектрограмма. Выбрана именно спектрограмма по мел-шкале, а не обычный её вид, так как в этом случае частоты, являющиеся наиболее значимыми для речи, становятся более различимыми на изображении для человеческого глаза, а значит и для методов анализа изображений. Сравнение типов спектрограмм можно увидеть на рисунке 1.

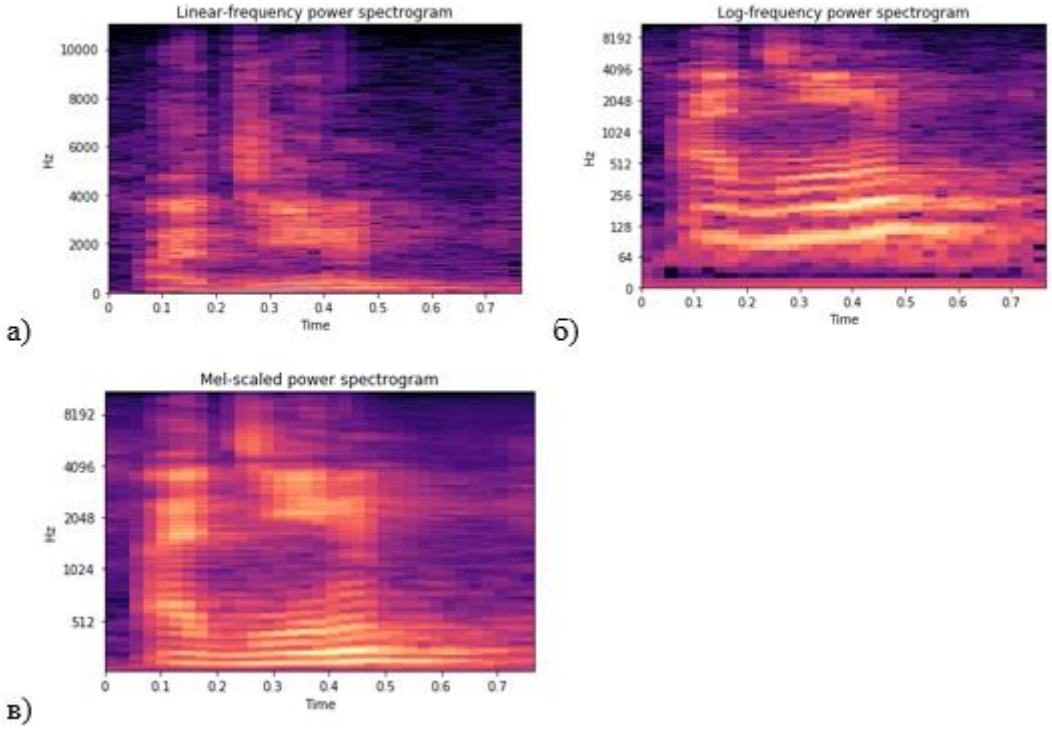

Рис. 1. Представление спектрограммы для слова *один* с шкалой *у* а) линейной б) логарифмической в) мел-шкалой.

Затем к каждой спектрограмме применяется ряд преобразований (рис. 2): нормализация для уменьшения влияния шума на изображение, фильтр Гаусса, а затем сжатие до матрицы размера 40x40.

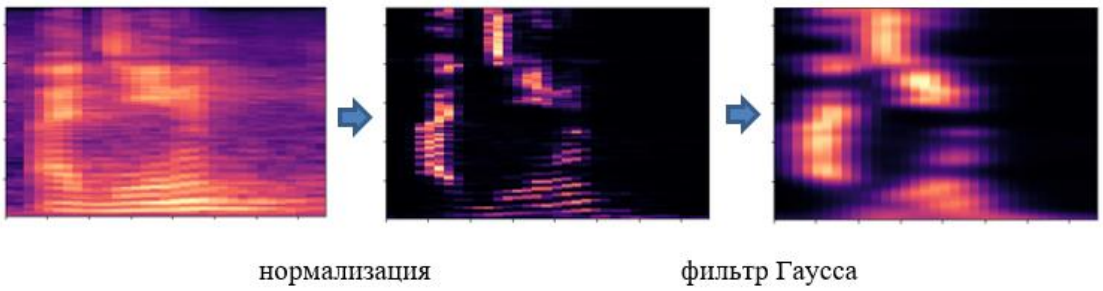

Рис. 2. Пример преобразований изображения.

Из преобразованного изображения создаётся вектор признаков с помощью метода создания гистограмм направленных градиентов (Histogram of Oriented Gradients). Для классификации полученных векторов используются методы RandomForestClassifier (с параметром n_estimators – 100) и SVM (с линейным типом ядра – linear).

5. Анализ результатов

В таблице 1 представлены точность результатов работы различных классификаторов для каждого способа извлечения признаков.

Табл. 1. Точность работы программы на тестовой выборке

	RandomForestClassifier	SVM	Neural Network
mfccs + chroma + mel + contrast	0.16	0.19	**0.36**
mfcc	0.15	**0.26**	0.02
mel-spectr + hog	**0.75**	0.59	0.05

Наибольшей точности классификации векторов, созданных при помощи вычисления средних значений четырёх популярных метрик, удалось добиться путём применения нейронной сети, но при этом точность составила всего 0.36, и при этом данный способ оказался наиболее затратным. Для обучения данной

модели понадобилось от 2,5 до 4 часов, в зависимости от конфигурации нейронной сети.

При классификации векторов, полученных из мел-кепстральных коэффициентов, наибольшую точность показал классификатор SVM, но точность составила 0.26. Результат превысил точность применения SVM на предыдущих векторах, и при этом потребовал небольших временных затрат, но всё же не является удовлетворительным.

Точность работы классификаторов на векторах, полученных в результате обработки спектрограмм, значительно превысил предыдущие методы. Лучший результат показал RandomForestClassifier – 0.75. Точность 75% можно считать достаточно значительной, учитывая небольшое количество и различающееся качество исходных данных.

Далее рассмотрим некоторые ошибки в работе итогового алгоритма. Подробную таблицу сравнения результатов наиболее точных вариантов алгоритмов можно найти в Приложении.

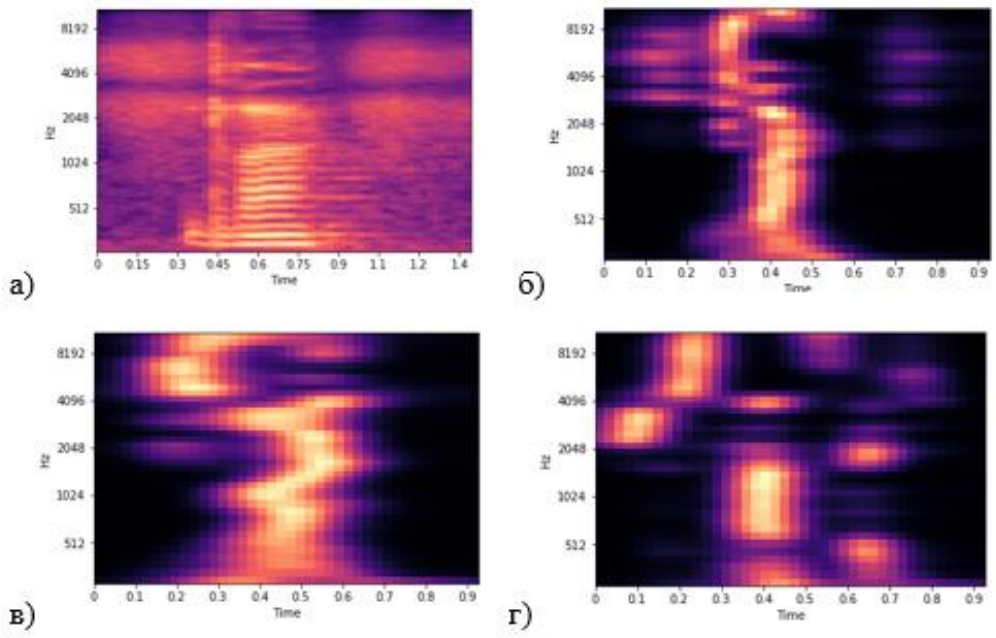

Рис. 3. Ошибочно распознанная спектрограмма для слова *два:* а) до обработки, б) после обработки; в) спектрограмма для слова *два* из обучающего набора; г) спектрограмма для слова *шестнадцать*.

Некоторые данные, судя по всему оказались слишком отличающимися от остальных по качеству. Например, один из экземпляров записи слова *два* не был верно размечен ни одним из сравниваемых алгоритмов. При рассмотрении соответствующей спектрограммы (рис. 3а) можно заметить, что в записи присутствует довольно сильный шум на высоких частотах, который не удаётся устранить после применения созданных фильтров (рис. 3б). При этом фрагменты записи с шумом не были обрезаны при автоматической сегментации данных, что привело к сжатию по горизонтали спектрограммы слова, что видно при сравнении с примером спектрограммы для слова *два* из обучающего набора данных (рис. 3в). В итоге изображению ошибочно было поставлено в соответствие слово *шестнадцать*, где также центральное место занимает фонема [а] (рис. 3г). Шум до и после слова *два* алгоритм был принят алгоритмом за первый и последний слоги слова *шестнадцать*.

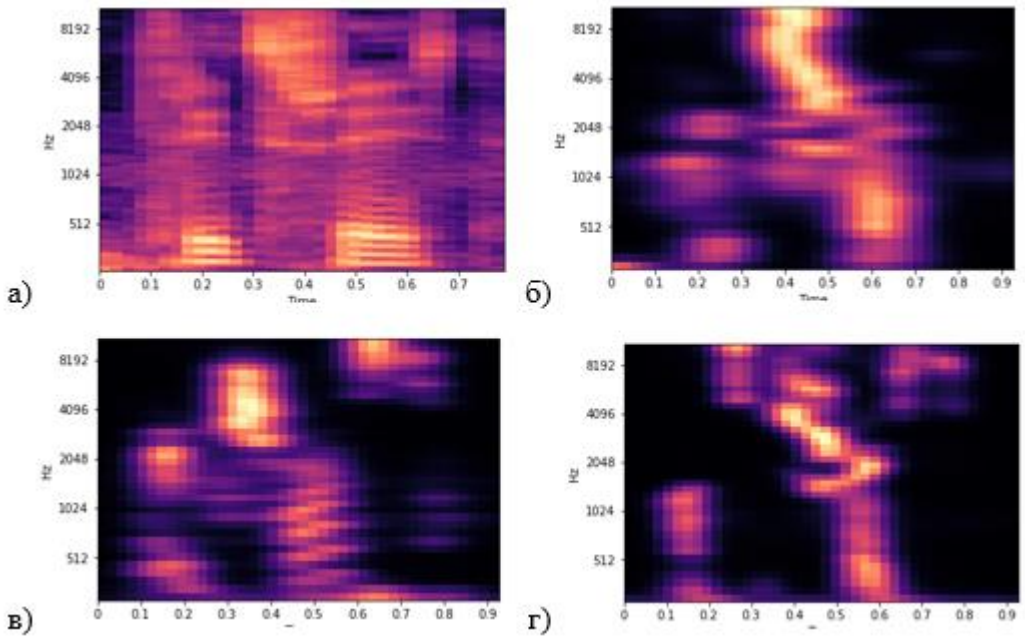

Рис. 4. Ошибочно распознанная спектрограмма для словосочетания *тридцать шесть:* а) до обработки, б) после обработки; в) спектрограмма для *тридцать шесть* из обучающего набора; г) спектрограмма для слова *двадцать шесть*.

Частым типом ошибок стали те случаи, когда вторая цифра в словосочетании распознавалась верно, а первая – ошибочно. Это можно объяснить тем, что вторая

цифро обычно произносилась с большим ударением и более чётко. Первая цифра, обозначающая десяток, произносилась тише и короче, при этом можно предположить, что такие слова, как, например, *двадцать* и *тридцать* будут недостаточно различимы. Например, эти факторы привели к ошибочному распознаванию словосочетания *тридцать шесть* (рис. 4а). При сравнении обработанных спектрограмм для *тридцать шесть* из тестового и обучающего наборов (рис. 4б и 4в) можно предположить, что на тестовой записи *тридцать* было произнесено недостаточно чётко для того, чтобы форманты фонемы [и] были разделимы и не сливались в единую форму на изображении, что скорее всего привело к ошибочному определению фразы, как *двадцать шесть* (рис. 4г).

6. Заключение

В ходе работы были собрано небольшое количество русскоязычных аудио данных для дальнейшей обработки. Проведено исследование методов классификации аудио данных, использующиеся в системах распознавания речи. Были реализованы варианты алгоритмов, не имеющие сложную структуру, основанные на популярных методах извлечения информации из аудио данных. Предложен и реализован подход к извлечению признаков для классификации из представления аудио в виде изображений спектрограмм. Предложенный метод имеет не сложную реализацию и показывает удовлетворительный результат (точность 75%) на ограниченном объёме разнообразных по свойствам русскоязычных данных.

Количество ошибок в работе алгоритма возможно значительно понизить с помощью дополнительной предобработки изначальных данных, для более качественной их сегментации и анализа. Но даже в предложенном виде модель покажет более качественный результат при увеличении объёма обучающих данных, как и любая обучаемая модель.

Результаты работы показывают, что предложенный метод сравним, если не превосходит по качеству простые реализации стандартных подходов, применимо к небольшим объёмам данных. Дальнейшее развитие методов автоматического анализа устной речи, основанных на анализе изображений, представляется крайне перспективным.

Литература

1. Schutte, K.T. (2009). *Parts-based Models and Local Features for Automatic Speech Recognition.*
2. Madan, A. & Gupta, D. (2014). *Speech Feature Extraction and Classification: A Comparative Review*. International Journal of Computer Applications (0975 – 8887) Volume 90 – No 9, March 2014
3. Dennis, J.W. (2014). *Sound Event Recognition in Unstructured Environments using Spectrogram Image Processing.*
4. Washani, N. & Sharma, S. (2015). *Speech Recognition System: A Review. International Journal of Computer Applications* (0975 – 8887) Volume 115 – No. 18, April 2015
5. Nguyen, Q.T. & Bui, T.D. (2016). *Speech classification using SIFT features on spectrogram images*. Vietnam Journal of Computer Science. November 2016, Volume 3, Issue 4, pp 247–257.

Приложение

Материалы в репозитории на Github:

https://github.com/SoDipole/ML-sound-vs-picture

Сравнение результатов работы алгоритмов:

SVM+combination			NN+combination			SVM + mfcc			RandForest + spectr		
true	result		true	result		true	result		true	result	
1	11		1	1	+	1	1	+	1	1	+
1	1	+	1	1	+	1	17		1	15	
2	2	+	2	2	+	2	2	+	2	2	+
2	45		2	40		2	44		2	16	
3	4		3	43		3	20		3	31	
3	23		3	3	+	3	3	+	3	3	+
4	3		4	3		4	34		4	4	+
4	3		4	4	+	4	3		4	4	+
5	45		5	45		5	5	+	5	5	+
5	13		5	13		5	2		5	5	+
6	6	+	6	6	+	6	9		6	6	+
6	6	+	6	46		6	9		6	6	+
7	9		7	33		7	4		7	7	+
7	9		7	24		7	7	+	7	7	+
8	38		8	8	+	8	8	+	8	8	+
8	8	+	8	8	+	8	8	+	8	8	+
9	24		9	24		9	8		9	9	+
9	9	+	9	9	+	9	10		9	9	+
10	30		10	30		10	9		10	10	+
10	30		10	30		10	30		10	10	+
11	15		11	11	+	11	35		11	11	+
11	31		11	31		11	13		11	30	
12	19		12	19		12	20		12	12	+
12	11		12	13		12	13		12	15	
13	15		13	12		13	18		13	12	
13	12		13	19		13	13	+	13	13	+
14	50		14	14	+	14	19		14	14	+
14	11		14	11		14	14	+	14	14	+
15	19		15	25		15	10		15	15	+
15	11		15	12		15	13		15	15	+
16	26		16	26		16	19		16	15	
16	17		16	17		16	11		16	16	+
17	15		17	15		17	18		17	14	
17	16		17	16		17	16		17	17	+
18	13		18	13		18	28		18	16	

18	18	+	18	18	+	18	29		18	18	+
19	8		19	41		19	44		19	19	+
19	11		19	11		19	14		19	17	
20	13		20	13		20	5		20	30	
20	20	+	20	20	+	20	12		20	20	+
21	27		21	21	+	21	33		21	21	+
21	27		21	27		21	21	+	21	23	
22	22	+	22	32		22	32		22	22	+
22	20		22	42		22	32		22	22	+
23	24		23	24		23	23	+	23	23	+
23	21		23	43		23	19		23	25	
24	4		24	14		24	27		24	24	+
24	38		24	35		24	41		24	24	+
25	21		25	25	+	25	35		25	25	+
25	13		25	25	+	25	27		25	25	+
26	26	+	26	26	+	26	26	+	26	26	+
26	46		26	26	+	26	26	+	26	26	+
27	21		27	28		27	24		27	27	+
27	29		27	29		27	24		27	27	+
28	18		28	18		28	28	+	28	28	+
28	27		28	27		28	11		28	23	
29	33		29	33		29	49		29	29	+
29	23		29	23		29	18		29	1	
30	30	+	30	30	+	30	16		30	30	+
30	30	+	30	7		30	3		30	5	
31	30		31	31	+	31	37		31	31	+
31	38		31	35		31	32		31	2	
32	22		32	25		32	29		32	32	+
32	22		32	32	+	32	22		32	32	+
33	9		33	3		33	33	+	33	33	+
33	9		33	9		33	33	+	33	31	
34	4		34	4		34	33		34	34	+
34	3		34	33		34	44		34	34	+
35	26		35	35	+	35	37		35	45	
35	14		35	35	+	35	36		35	22	
36	36	+	36	36	+	36	4		36	36	+
36	36	+	36	7		36	10		36	26	
37	37	+	37	39		37	37	+	37	37	+
37	33		37	10		37	31		37	37	+
38	29		38	38	+	38	28		38	38	+
38	38	+	38	37		38	34		38	38	+
39	39	+	39	39	+	39	38		39	39	+
39	14		39	39	+	39	36		39	35	
40	20		40	48		40	7		40	40	+

40	20		40	40	+	40	16		40	40	+
41	20		41	41	+	41	41	+	41	41	+
41	49		41	8		41	37		41	41	+
42	40		42	42	+	42	42	+	42	32	
42	32		42	13		42	42	+	42	42	+
43	24		43	44		43	48		43	43	+
43	49		43	43	+	43	45		43	42	
44	35		44	24		44	49		44	44	+
44	6		44	43		44	49		44	44	+
45	45	+	45	45	+	45	45	+	45	45	+
45	25		45	6		45	48		45	45	+
46	6		46	46	+	46	46	+	46	46	+
46	26		46	46	+	46	36		46	46	+
47	10		47	49		47	34		47	47	+
47	4		47	44		47	46		47	47	+
48	28		48	28		48	48	+	48	48	+
48	20		48	28		48	48	+	48	48	+
49	35		49	30		49	49	+	49	49	+
49	29		49	49	+	49	16		49	45	
50	50	+	50	50	+	50	50	+	50	50	+
50	28		50	21		50	8		50	50	+