УДК 534.4+004.93

ПРИГОДНОСТЬ РЕЧЕВЫХ ФАЙЛОВ ДЛЯ СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ ПОСЛЕ ПРОЦЕДУРЫ ОЧИСТКИ ОТ ШУМА

Р.Х. Латыпов, Е.Л. Столов

Аннотация

Приведены результаты экспериментов по очистке речевых файлов от шума с последующей пересылкой файла в систему распознавания речи Google. Предложен оригинальный метод обнаружения интервалов в файле, содержащих только шум. Показано, что использование модифицированного фильтра Винера во временной области позволяет улучшить качество распознавания зашумленных файлов.

Ключевые слова: речевой файл, очистка от шума, распознавание речи, Google, фильтр Винера.

Введение

Проблема автоматического распознавания речи перешла из области теории и опытных реализаций в реальную жизнь. Наиболее яркой демонстрацией этого явления стала система обработки речевых запросов в Google. Другим примером может служить система синхронного перевода для Skype, предложенная MicroSoft. Несмотря на огромные успехи, достигнутые в указанном направлении, известные системы страдают существенным недостатком: даже небольшое присутствие шума в речевом сигнале, которое не доставляет проблем человеку для восприятия речи, оказывается фатальным препятствием для систем автоматического распознавания речи (САРР). С другой стороны, в настоящее время разработано огромное количество методов очистки речевых файлов от шума [1, 2]. Последние результаты, полученные в данной области, представлены в [3]. При оценке качества предлагаемых алгоритмов применяются различные методы. Из них наиболее надежным, но и более дорогим с точки зрения затраченных ресурсов является представление результатов обработки файла человеческой аудитории. Слушателям раздается специальная анкета, в которой они должны отметить слова или предложения, воспринятые в результате прослушивания. Поскольку организация подобной проверки встречает определенные трудности, разработаны способы автоматической оценки качества восстановленного файла: SNR, PESQ [1]. В то же время подача на вход САРР восстановленного файла приводит, как правило, к отказу системы от распознавания либо к неверному результату. Это связано с тем, что в результате работы алгоритма восстановления часто осуществляется синтез речевого сигнала, а САРР ориентирована на естественную речь. В этой связи возникла задача оценки возможности сохранить файл пригодным для САРР в результате работы различных алгоритмов очистки речевых файлов от шума. В настоящее время проблема оценки качества восстановленного сигнала существенно упростилась благодаря доступу к API для распознавания речи, представленному корпорацией Google. Именно этот метод реализован в настоящей статье.

В п. 1 дается краткий обзор известных методов очистки файла от шума, в п. 2 описывается оригинальный подход для выделения фрагментов, содержащих шум,

в п. 3 предлагается модификация фильтра Винера во временной области, учитывающая конечность обрабатываемых фрагментов, в п. 4. представлены результаты экспериментов.

1. Обзор методов очистки речевых файлов от шума

Все методы очистки сигнала, рассматриваемые в работе, базируются на аддитивной модели (1):

$$y(t) = x(t) + n(t). \tag{1}$$

Здесь $y(t),\ x(t),\ n(t)$ – наблюдаемый сигнал, чистый сигнал и шум соответственно. Шум и чистый сигнал не коррелируют, причем относительно шума делаются различные предположения. Речевой файл разбивается на пересекающиеся фреймы длиной 10–20 мс, внутри которых сигнал считается стационарным, и к нему применяются известные методы обработки стационарных сигналов. Алгоритмы удаления шума разбиваются на два класса: для приема сигнала используется один микрофон либо несколько микрофонов. В статье рассматривается только первый случай.

Все методы, применяемые для подавления шума, можно весьма условно разбить на следующие группы:

- преобразования во временной области;
- преобразования в частотной области;
- проектирование на подпространство на основе предположений о размерности пространства речевых сигналов;
- использование кодовой книги для исправления искаженных участков речи.

Охарактеризуем каждую из этих групп.

Преобразование во временной области. К речевому сигналу применяется линейный фильтр, спроектированный таким образом, чтобы подавить шум. Для этой ситуации подходит фильтр Винера, если точно известны параметры шума. На практике эти параметры приходится извлекать из наблюдаемого сигнала либо делать априорные предположения. Модификация данного метода рассмотрена ниже.

Преобразование в частотной области. Это наиболее популярный метод, применяемый в настоящее время. Весьма полный обзор модификаций используемых алгоритмов представлен в [3]. В основе алгоритмов лежит предположение, основанное на физиологии восприятия речи человеком, согласно которому фаза сигнала не влияет на разборчивость речи. В этой связи все внимание обращено на модификацию модулей коэффициентов Фурье с целью убрать компоненты шума. Это направление получило название «спектральное вычитание», фактически это есть дальнейшее развитие метода, предложенного в [4]. Чистый сигнал и шум считаются случайными величинами с нулевым средним. Применим к обеим частям равенства (1) дискретное преобразование Фурье. Поскольку сигнал и шум не коррелируют, это также справедливо и для их коэффициентов Фурье, которые тоже имеют нулевые средние. Используя утверждение о дисперсии суммы некоррелированных сигналов, получим

$$\sigma^2(|Y(m)|) = \sigma^2(|X(m)|) + \sigma^2(|N(m)|)$$

для каждого индекса m коэффициента Фурье. Зная спектр шума, получаем простейшую оценку модуля чистого сигнала, приравнивая модуль сигнала среднеквадратическому отклонению:

$$|X(m)| = \sqrt{|Y(m)|^2 - \sigma^2(|N(m)|)}.$$

При практическом использовании этого метода пользуются оценкой вида

$$|X(m)| = \sqrt{|Y(m)|^2 - \mathrm{coeff} \cdot \sigma^2(|N(m)|)}.$$

Здесь coeff > 1 – коэффициент, который подбирается опытным путем. Если выражение под корнем становится отрицательным, то $X(m)$ заменяется нулем, в остальных случаях полагают $X(m) = |X(m)| \exp(2\pi i w)$, где w – либо случайная величина, равномерно распределенная на интервале [0,1], либо фаза $Y(m)$. С помощью обратного преобразования Фурье восстанавливается чистый фрейм. Восстановленные таким образом фреймы склеиваются в выходной сигнал.

Направление дальнейших исследований сводится к уточнению оценок $|X(m)|$. Первоначально были предложены параметрические модели распределения модуля [5, 6]. По измеренным значениям дисперсий уточнялись параметры модели, после чего подсчитывалось наиболее вероятное значение модуля. Этот подход требовал значительных вычислений, но не приводил к существенному улучшению качества. Дело в том, что предположения модели подходили не для всех фреймов, в частности, они не годились для фреймов, содержащих согласные. В последнее время стали более популярными методы оценки $|X(m)|$ на основе значений дисперсии в предыдущих фреймах. Оценка в текущем фрейме строилась в виде линейной комбинации значений в предыдущих фреймах. Эта техника не предполагает известным вид распределения, и задача сводится к способам выбора коэффициентов линейной комбинации [3]. Недостатком указанного подходя является случайный выбор фазы сигнала, в результате чего файл становится непригодным для САРР. Некоторое улучшение ситуации достигается путем замены преобразования Фурье дискретным косинус-преобразованием. В этом случае все коэффициенты являются вещественными числами, и вместо фазы комплексного числа надо выбрать только знак вещественного числа. Как и выше, этот знак полагается равным знаку $Y(m)$. При таком подходе, хотя и повышается пригодность восстановленного файла для САРР, количество правильно распознанных слов остается недостаточным.

Проектирование на подпространство. Рассмотрим часть файла, лежащую внутри фрейма длины M. Можем считать все такие части векторами пространства размерности M. Метод основан на предположении, что векторы, принадлежащие чистому сигналу и шуму, лежат в разных линейных подпространствах. Строятся преобразования проектирования на эти подпространства, и таким образом получают очищенный сигнал. Искомые подпространства конструируют на основе метода главных компонент, при этом приходится решать большое количество задач на нахождение собственных векторов симметрической матрицы порядка M (см. [2]). Методу присущи следующие недостатки. Предположение о том, что проекции исходного сигнала на найденные подпространства имеют в пересечении векторы, близкие к нулевому вектору, трудно обосновать. Кроме того, не существует надежного метода априорного определения размерностей нужных подпространств.

Применение кодовой книги. Наиболее перспективной представляется технология, основанная на применении кодовой книги [7, 8]. В этом случае создается кодовая книга, содержащая набор чистых фонем. Система предварительно очищает файл от шума одним из приведенных выше способов, после чего синтезируется новый файл на основе сравнений модулей преобразования Фурье очищенного файла и спектра фонем из кодовой книги. Авторы пишут о высоком качестве полученного сигнала, но сам метод требует больших вычислений.

Все рассмотренные выше методы предполагают знание параметров шума. Внутри фрейма шум считается стационарным, но он может меняться от фрейма к фрейму. В этой связи предлагается определять параметры шума по участкам между фразами или между словами, где отсутствует речевой сигнал. Интерес представляют методы, позволяющие производить такое выделение в автоматическом режиме. В настоящее время предложено большое количество подходов для решения указанной задачи, обзор которых можно найти в [9]. В их основе лежит

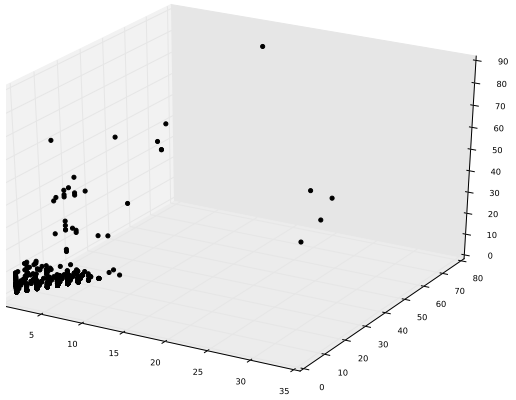

Рис. 1. Распределение признаков фреймов по трем коэффициентам

идея оценки изменения мощности сигнала. В простейшем варианте ставится порог, и фрейм, в котором мощность меньше заданного порога, объявляется шумом. В расширенном варианте исследуется распределение мощности в разных полосах спектра, и решение принимается на основе этого распределения.

2. Выделение шумовых участков

В настоящей работе предлагается оригинальная вычислительная процедура для выделения фреймов с шумом, которая хорошо согласуется с наблюдениями и не базируется на оценке мощности сигнала. К каждому фрейму после сглаживания применяют преобразование Фурье и получают локальный спектр. Из полученного набора коэффициентов Фурье выбирают заданное число $NumPos$ коэффициентов, имеющих наибольшие модули, и сохраняют их номера. Этот набор номеров назовём признаком фрейма. Гипотеза заключается в том, что в случае стационарного шума эти признаки мало меняются при переходе от одного фрейма с шумом к другому аналогичному фрейму. На рис. 1, построенном с помощью функций библиотеки Matplotlib [11], представлен пример распределения признаков фрейма из трех номеров коэффициентов. По осям координат отложены номера, а сам признак обозначен точкой. Как следует из этого рисунка, наблюдается скопление признаков в одном месте. Для отыскания точки сгущения применяется стандартная процедура:

- находим вектор, полученный усреднением всех найденных векторов-признаков, и округляем полученные значения;
- в цикле отбрасываем наиболее удаленные от центра векторы;
- вычисляем центр оставшегося множества множества и округляем полученные значения;
- выход из цикла осуществляется после того, как центр не изменяется после очередного шага.

После того как определен центр множества, в качестве фрейма с шумом объявляется фрагмент, чей признак находится от центра не более чем на заданном расстоянии.

Предварительная обработка файла сводится к следующей процедуре:

- просматриваются фреймы заданного размера с перекрытием на половину длины и для каждого из них подсчитывается преобразование Фурье;
- подсчитывается признак фрейма, состоящий из $NumPos$ чисел, и сравнивается с значениями центра, вычисленными ранее;

• если найденный фрейм помечается как содержащий шум, то он используется в процедуре очистки для всех последующих фреймов до тех пор, пока не будет найден следующий фрейм, содержащий шум;

• к остальным фреймам применяется процедура очистки с использованием шума, найденного ранее, после чего очищенные фреймы склеиваются и образуют выходной файл.

3. Модификация фильтра Винера для очистки сигнала от шума

После того как найдена составляющая $n(t)$ в (1), применяют один из подходов для очистки шума, приведенных выше. Поскольку ставится задача сделать файл пригодным для САРР, принято решение использовать алгоритм, наименее «травмирующий» спектр сигнала. Таковым является фильтр Винера во временной области. При стандартном подходе (см., например, [10]) в каждом фрейме длины M подсчитываются автокорреляционные функции $R_y(p)$ и $R_n(p)$ наблюдаемого сигнала и шума соответственно, находится оценка автокорреляционной функции исходного сигнала

$$R_x(p) = R_y(p) - R_n(p) \tag{2}$$

и решается система уравнений

$$\sum_{m=0}^{M-1} R_y(m-i)b_m = R_x(i), \quad i = 0, \ldots, M-1. \tag{3}$$

Если найдены коэффициенты b_m, то результаты $\widehat{x}(k)$ фильтрации определяются по формуле

$$\widehat{x}(k) = \sum_{m=0}^{M-1} b_m y(k-m). \tag{4}$$

При обосновании данного алгоритма возникают следующие проблемы: в оригинальной теории сумма в (4) является бесконечной, поэтому неясно, как надо выбирать M; как оценивать значения автокорреляционных функций по конечному фрейму, когда M сравнимо с длиной фрейма. Более того, для разных значений i фактическое число слагаемых в (4) оказывается разным. Для совпадения длин исходного фрейма и фрейма, полученного после фильтрации, приходится добавлять нули в исходный фрейм. Все это осложняет прямое применение данной процедуры. В этой связи предлагается модификация алгоритма, в которой указанные проблемы формально снимаются. Выбирается число W – количество коэффициентов в фильтре Винера – таким образом, чтобы $M - W + 1$ было четным числом, и рассматривается система линейных уравнений

$$\mathbf{F} \cdot \mathbf{A} = \mathbf{X}, \tag{5}$$

где

$$\mathbf{F} = \begin{pmatrix} y_0 & y_1 & y_2 & \cdots & y_{W-1} \\ y_1 & y_2 & y_3 & \cdots & y_W \\ y_2 & y_3 & y_4 & \cdots & y_{W+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{M-W} & y_{M-W+1} & y_{M-W+2} & \cdots & y_M \end{pmatrix},$$

$$\mathbf{A} = (a_0, a_1, \ldots, a_{W-1})^T, \quad \mathbf{X} = (x_0, x_1, x_2, \ldots, x_{M-W})^T.$$

Для отыскания псевдорешения $\mathbf{A}$ системы (5), как обычно, перейдем к уравнению

$$\mathbf{F}^T \cdot \mathbf{F} \cdot \mathbf{A} = \mathbf{F}^T \cdot \mathbf{X},$$

Табл. 1

Результаты распознавания до и после очистки файлов

Число слов	До очистки		После очистки	
	найдено	число ошибок	найдено	число ошибок
10	5	0	6	0
7	4	3	5	1
7	6	1	6	1
7	6	1	6	0
7	7	1	7	1
8	8	5	7	1
8	3	2	0	0
6	6	0	6	0
6	4	2	6	1
6	6	2	6	1
8	7	0	8	0
8	7	0	8	0

после чего заменим вектор $\mathbf{F}^T \cdot \mathbf{X}$ согласно (2) его оценкой $\mathbf{R_y} - \mathbf{R_n}$. На практике, как и в случае спектрального вычитания, лучший результат получается при замене указанного выражения на $\mathbf{R_y} - \text{coeff} \cdot \mathbf{R_n}$, где coeff подбирается экспериментально из интервала $0.4 \div 0.8$. Решая полученную систему, находим оценку вектора $\mathbf{A}$, затем из системы (5) находим оценки $M - W + 1$ значений чистого сигнала $\mathbf{X}$. Поскольку указанное количество сигналов является четным, устанавливаем шаг перекрытия фрагментов равным половине этого значения, что дает удобную возможность склеить все найденные фрагменты в один выходной сигнал.

4. Результаты эксперимента

Для проверки работоспособности предложенной процедуры была выбрана небольшая база, предоставленная Московским отделением Samsung. База состоит из 12 файлов, записанных с частотой 22050 Гц. Каждый файл содержал единственное предложение и был искажен аддитивным шумом с SNR порядка 3 дБ. Зашумленный файл воспринимался человеком без особых затруднений. С помощью программы на основе API Google каждый файл был послан для распознавания до и после обработки. В качестве ответа программа выдавала новое предложение. Для оценки качества распознавания были использованы два значения: разность между числом слов в исходном и выведенном предложениях и число ошибок в выведенном предложении. Длина фрейма выбиралась в интервале $20 \div 30$ мс, а количество коэффициентов W – из интервала $10 \div 20$. Полученные данные приведены в табл. 1.

В левом столбце указано число слов в оригинальном предложении, в оставшихся столбцах приведены количество слов в распознанной фразе и число неверных слов.

Выводы

Приведенные результаты экспериментов показывают, что предложенная техника улучшает качество распознавания. Следует отметить, что использованный распознаватель от Google анализирует предложение целиком, а не по отдельным словам. В результате иногда возникают «фантастические» варианты, почерпнутые из каких-то источников и не имеющие отношения к оригинальным предложениям. Заметим также, что очищенные файлы воспринимаются на слух гораздо лучше, чем зашумленные оригиналы.

Работа выполнена при частичной финансовой поддержке Управления высокопроизводительных алгоритмов в Исследовательском Центре Самсунг в Москве.

Summary

R.Kh. Latypov, E.L. Stolov. Suitability of Speech Files for Automatic Speech Recognition Systems after Noise Reduction Procedures.

The results of the experiments on noise reduction in speech files with further transfer in the Google automatic speech recognition (ASR) system are presented. An original method is developed to identify the intervals containing only noise. It is shown that using a modified Wiener filter in the time domain allows to improve the recognition quality of noisy files.

Keywords: speech file, noise reduction, speech recognition, Google, Wiener filter.

Литература

1. *Quatieri T.F.* Discrete-Time Speech Signal Processing. – Prentice-Hall PTR, 2002. – 781 p.
2. *Benesty J., Chen J., Huang Y., Cohen I.* Noise Reduction in Speech Processing. – Berlin; Heidelberg: Springer-Verlag, 2009. – 240 p.
3. *Hendriks R.C., Gerkmann T., Jensen G.J.* DFT-Domain Based Single Microphone Noise Reduction for Speech Enhancement. – Morgan & Claypool Publ., 2013. – 80 p.
4. *Schroeder M.R.* Apparatus for suppressing noise and distortion in communication signals: U.S. Patent No. 3,180,936, filed Dec. 1, 1960, issued Apr. 27, 1965.
5. *McAulay R.J., Malpass M.L.* Speech enhancement using a soft-decision noise // EEE Trans Acoust., Speech, Signal Processing. – 1980 – V. ASSP-28, No 2. – P. 137–145.
6. *Ephraim Y., Malah D.* Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator // IEEE Trans Acoust., Speech, Signal Processing. – 1984. – V. ASSP-32, No 6. – P. 1109–1121.
7. *Srinivasan S., Samuelsson J., Kleijn W.B.* Codebook-Based Bayesian Speech Enhancement for Nonstationary Environments // IEEE Trans. Audio, Speech, Language Processing. – 2007. – V. 15, No 2. – P. 441–452.
8. *Rozenkranz T.* Modeling the temporal evolution of LPC parameters for codebook-based speech enhancement // Proc. 6th Int. Symposium on Image and Signal Proc. and Analysis. – IEEE, 2009. – P. 455–460.
9. *Marzinzik M., Kollmeier B.* Speech pause detection for noise spectrum estimation by tracking power envelope dynamics // IEEE Trans. Speech and Audio Processing. – 2007. – V. 10, No 2. – P. 109–118.
10. *Chen J., Benesty J., Yiteng Huang, Doclo S.* New insights into the noise reduction Wiener filter // IEEE Trans. Audio, Speech, Language Processing. – 2006. – V. 14, No 4. – P. 1218–1234.
11. *Hunter J.D.* Matplotlib: A 2D Graphics Environment // Computing in Science & Engineering. – 2007. – V. 9, No 3. – P. 90–95.

Поступила в редакцию
26.08.15

Латыпов Рустам Хафизович – доктор технических наук, директор Института вычислительной математики и информационных технологий, Казанский (Приволжский) федеральный университет, г. Казань, Россия.

E-mail: *Roustam.Latypov@kpfu.ru*

Столов Евгений Львович – доктор технических наук, профессор кафедры системного анализа и информационных технологий, Казанский (Приволжский) федеральный университет, г. Казань, Россия.

E-mail: *ystolov@kpfu.ru*