МЕТОД СЕГМЕНТАЦИИ СПЕКТРОГРАММ РЕЧЕВОГО СИГНАЛА

Литвиненко С.Л.
Волгоградский государственный технический университет

Задача сегментации речевого сигнала является очень важной в теории распознавания речи. Она решается как при создании обучающих баз данных, содержащих фразы с информацией об их сегментации на звуки, так и во время работы систем распознавания слитной речи, основанных на фонемном подходе для выделения из речевого потока конкретных звуков.

Для сегментации речевого сигнала применяются так называемые методы детектирования изменений сигнала либо методы кластеризации данных. Любой метод детектирования изменений основан на следующих основных действиях: моделирование сигнала; вычисление ошибки предсказания исходного сигнала по построенной модели; принятие решения по функции ошибки о наличие изменения в сигнале. По этой функции сигнал может быть разбит на сегменты, внутри которых остаются постоянными в статистическом смысле и отличными от других сегментов некоторые характеристики данного сигнала.

В алгоритмах детектирования изменений сигнал оценивается с использованием одной модели (фильтра), двух или нескольких моделей [1]. При использовании двух моделей сравниваются ошибки предсказания сигнала каждой моделью. Параметры одной модели оцениваются по сигналу на коротком отрезке времени, а параметры другой – на более длинном отрезке времени. Изменение считается детектированным, если ошибка с выхода модели, настроенной по большему сегменту, будет больше чем с выхода модели, настроенной по меньшему сегменту данных. В моногомодельном подходе используются априорные сведения о свойствах сигнала. По этим данным строятся согласованные фильтры, настроенные на известные заранее характеристики сигнала. Таким образом, каждый фильтр реагирует на свой сегмент сигнала.

В последнем случае для детектирования изменений решается задача классификации сигнала, так как каждый сегмент сигнала можно рассматривать как данные, относящиеся к одному определенному классу. На таком же подходе основываются алгоритмы сегментации речевого сигнала выполняющие его кластеризацию [3]. В этом случае сигнал рассматривается как набор векторов в признаковом пространстве. Алгоритм кластеризации производит поиск групп векторов наиболее близко расположенных друг к другу, определяя для них некоторый усредненный вектор – центр кластера.

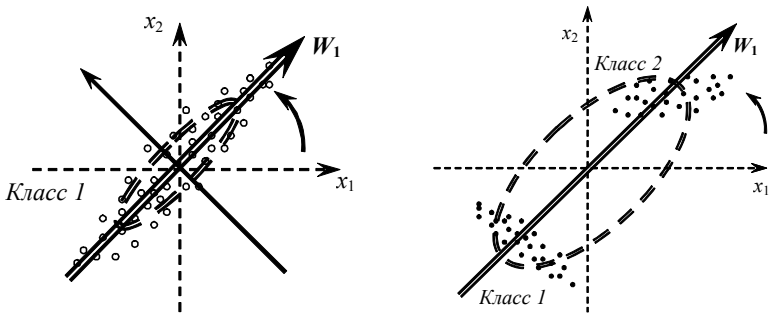
Итак, задача сегментации сигнала тесно связана с задачей его классификации. Наилучшие результаты работы показывают алгоритмы сегментации, использующие для своей работы классификацию сигнала (многомодельный подход в детектировании переходов и методы кластеризации). Однако эти алгоритмы не используют информацию об изменении характеристик сигнала при переходе между сегментами, которую можно получить в двухмодельном подходе, анализируя всего лишь два соседних сегмента сигнала.

Предлагаемый в данной работе метод сегментации сочетает в себе одновременно два подхода, с одной стороны сигнал рассматривается как набор векторов признаков, которые можно классифицировать, с другой стороны окончательная классификация сегментов не производится, а вычисляется только функция ошибки с выхода модели как в первых двух методах детектирования переходов. Сигнал моделируется с помощью одной модели, а ошибка принимает большие значения, когда в моделируемый сегмент сигнала попадают вектора признаков из двух различных классов. То есть, в терминах методов классификации, данный алгоритм определяет моменты времени, в которые происходит смена классов векторов признаков исследуемого сигнала.

Метод основан на анализе речевого сигнала представленного в виде его спектра мощности. Спектр мощности, полученный, например, с использованием кратковременного преобразования Фурье, будем рассматривать как отображение дискретной функции одной переменной, задающей изменение энергии речевого сигнала во времени, в многомерное пространство признаков.

Метод основан на допущении, что в пределах одного класса, в простейшем случае - звука речи, постоянными характеристиками на протяжении одного звука речи являются положение его спектральных максимумов. А это определенный набор компонент (координат) векторов анализируемого сегмента, которые на протяжении данного сегмента изменяются синхронно и гораздо сильнее остальных компонент имеющих меньшую величину. Тогда в первом приближении весь звук можно охарактеризовать некоторым характерным вектором и функцией изменения его по мощности, что будет соответствовать изменению мощности самых больших компонент вектора. Эти две характеристики и входят в модель, с помощью которой сигнал описывается на протяжении одного звука.

Если мы будем рассматривать некоторый набор векторов $\boldsymbol{Y}=\{\boldsymbol{y}_1, \boldsymbol{y}_2,\ldots \boldsymbol{y}_m\}$ расположенных в одном окне анализа относящихся только к одному классу, то в пространстве признаков они будут располагаться близко в виде некоторого облака, которое для двумерного случая будет выглядеть как показано на рис. 1. Так как мы приняли, что лишь синхронное изменение некоторых компонент вектора для нас имеет значение, то это облако будет больше всего вытянуто в пространстве лишь вдоль одного направления. Это направление, определяемое дисперсией вдоль осей $\boldsymbol{x}_1$ и $\boldsymbol{y}_2$, показанное на рисунке двойной стрелкой, необходимо найти при построении модели для конкретного звука.



а) **б)**

Рис. 1. Два варианта расположения первого главного собственного вектора, (а) - если анализируются данные одного класса, (б) – двух классов

Для определения такого направления, в котором дисперсия данных наибольшая, применим известный в статистике метод анализа главных компонент, называемый также преобразованием Хоттелинга, а в теории цифровой обработки сигналов – преобразованием Карунена-Лоэва [2], [3], которое представляет собой линейное ортогональное преобразование вида $Y = W \cdot X$.

Оно трансформирует вектор $\boldsymbol{X}$ в вектор $\boldsymbol{Y}$ посредством матрицы $\boldsymbol{W}$. Для определения матрицы $\boldsymbol{W}$ рассматривается набор векторов $\boldsymbol{X}$, т.е. матрица, состоящая из этих векторов. Для нее вычисляется ковариационная матрица

$$\Sigma_X = E\{(X - \overline{X})(X - \overline{X})^T\}$$

Матрица $\boldsymbol{W}$ определяется таким образом, чтобы составляющие ее вектора $\boldsymbol{w_i}$ были собственными векторами ковариационной матрицы $\sum_X$, тогда ковариационная матрица в области изображений

$$\Sigma_Y = E\{(Y - \overline{Y})(Y - \overline{Y})^T\} = W\Sigma_X W^{-1} = W\Sigma_X W^T$$

будет диагональной, состоящей из собственных значений матрицы $\sum_X$ равных $\Sigma_Y = diag(\lambda_1, \lambda_1, \ldots, \lambda_N)_i$, где λ_i – собственные значения $\sum_X$, и $\Sigma_X w_i = \lambda_i w_i$. Но на диагонали ковариационной матрицы находятся дисперсии компонент векторов матрицы $\boldsymbol{Y}$. Значит $\boldsymbol{\lambda_i}$ – представляют собой дисперсии компонент векторов матрицы $\boldsymbol{Y}$. Тогда, так как матрица $\sum_Y$ диагональная, компоненты векторов матрицы $\boldsymbol{Y}$ некоррелированы. Исходная матрица $\boldsymbol{X}$ может быть восстановлена только по тем компонентам $\boldsymbol{Y}$, которые имеют наибольшие дисперсии равные $\boldsymbol{\lambda_i}$, а значит, несут наибольшую информацию об исходном сигнале: $\widetilde{X} = W'^T Y'$

Алгоритмы нахождения собственных значений выдают их упорядоченными по возрастанию. Сохранив только те компоненты $\boldsymbol{Y}$ и $\boldsymbol{W}$, которые соответствуют первым $\boldsymbol{M}$ собственным числам можно сжимать информацию, представленную сигналом $\boldsymbol{X}$.

Графически данное преобразование можно представить как вращение системы координат, связанной с векторами данных таким образом, чтобы оси расположились вдоль векторов матрицы $\boldsymbol{W}$, называемых главными векторами. Первая из осей будет располагаться по направлению наибольшей вариации данных, а последняя – наименьшей (см. рис. 1 а.).

Из указанного ранее предположения следует, что при описании одного звука речи достаточно рассматривать в его разложении по главным компонентам только первую главную компоненту $\boldsymbol{Y_1}$. Тогда эта компонента будет характеризовать изменение мощности сигнала, а первый главный собственный вектор $\boldsymbol{W_1}$, соответствующий первой компоненте, задающий координаты направления по которому вариация данных наибольшая, соответствует характерному вектору для данного класса.

Теперь рассмотрим вектора, принадлежащие двум разным классам. Облака, соответствующие этим векторам схематически изображены на рис. 1 б). Если мы попытаемся описать два класса, то есть два смежных звука речи, то ось первой главной компоненты будет направлена примерно по линии, соединяющей центры классов (если классы расположены друг от друга на расстоянии большем, чем внутриклассовые расстояния между векторами каждого класса). Естественно, что в общем случае, ось первой главной компоненты не будет характеризовать ни средний вектор первого класса, ни средний вектор второго класса. Поэтому спектрограмму двух смежных звуков речи по одной компоненте восстановить не удастся, появится ошибка восстановления гораздо большая, чем ошибка восстановления всего лишь одного звука только по одной компоненте. На измерении этой ошибки и основан данный метод сегментации спектрограммы речевого сигнала.

При реализации данного алгоритма не требуется рассчитывать все главные компоненты - собственные вектора ковариационной матрицы, необходим только один, соответствующий наибольшему собственному значению. Традиционные же итерационные алгоритмы, использующие для этого QR-разложение, вычисля-

ют сразу все собственные значения. Поэтому лучше применять нейросетевой подход, предложенный Е. Ойа (см. [4]).

Для определения первого главного компонента Y_1 и главного собственного вектора W_1, связанного с ним, будем использовать один линейный нейрон, на вход которого подаются вектора X_i из матрицы данных X, а на выходе получается значение коэффициента Y_{1i}: $Y_{1i} = W_1^T X_i = \sum_{j=1}^{N} W_{1j} X_{ij}$

Вектор W_1 весов нейрона, сходящийся после обучения нейрона к первому главному собственному вектору данных вычисляется следующим образом: $W_1(t+1) = W_1(t) + sY_1(X_i(t) - W_1(t)Y_1(t))$

Где коэффициент s задает скорость обучения. Итак, для нахождения первого главного собственного вектора нейрон обучается на данных X из окна анализа. Затем производится восстановление исходного сигнала $\tilde{X}_i = W_1 Y_{1i}$ и вычисляется ошибка восстановления: $E_k = \sum_i \left(X_i - \tilde{X}_i\right)^2$,

Где индекс k обозначает номер текущего окна, по которому производится оценка ошибки восстановления спектрограммы. В полученной таким образом функции ошибки ищется положение локальных максимумов, соответствующих переходам между звуками.

Для нахождения функции ошибки описания сигнала нашей моделью спектрограмма просматривается через смещающееся по оси времени окно. Далее выполняется разложение с использованием описанного выше нейрона по первой главной компоненте фрагмента спектра, умещающегося в таком окне. Затем восстанавливается исходная спектрограмма из этого окна, производится вычисление ошибки восстановления. Указанная процедура выполняется для всех положений данного окна во времени, в результате чего получается функция ошибки восстановления исходного сигнала. Эта функция имеет локальные минимумы в те моменты, когда окно находится в пределах только одного звука, и локальные максимумы, когда окно находится на границе между двумя звуками. Причем, чем сильнее звуки отличаются друг от друга (мера расстояния Евклидова), тем сильнее будет скачок функции ошибки. Программа сегментации выделяет локальные максимумы функции ошибки, и размечает речевой поток на основании полученной информации. Результат работы алгоритма поиска переходов показан на рис. 2. Моделирование проводилось в среде MATLAB, речевой сигнал оцифровывался с частотой 16кГц, в качестве сегментируемых слов были выбраны 10 цифр, произнесенных одним диктором. Размер окна при вычислении кратковременного Фурье преобразования был выбран равным N = 256 выборок, шаг смещении k = 100 выборкам. После вычисления амплитудного спектра спектрограмма сглаживалась двумерным фильтром скользящего среднего с размером окна 4 выборки по шкале частот и 6 выборок по оси времени, функция ошибки предсказания вычислялась по окну размером 6 выборок по оси времени спектрограммы, данное значение было выбрано экспериментально как компромисс между наличием в функции ошибки ложных максимумов и отсутствием некоторых из них. Для данных из каждого окна выполнялось 15 циклов обучения нейрона со скоростью обучения s = 0.0001.

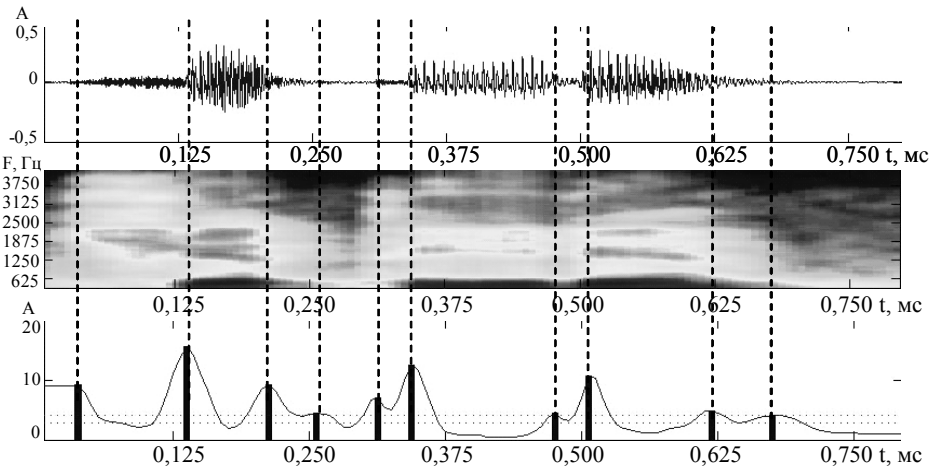


Рис. 2. Иллюстрация работы алгоритма сегментации речевого сигнала. Вверху - график сигнала, посередине – спектрограмма сигнала, внизу график функции ошибки восстановления сигнала по первой главной компоненте, жирными линиями отмечены моменты переходов от одного звука к другому, пунктиром выделены сегменты сигнала

Функцию ошибки восстановления (см. рис. 2.) можно также интерпретировать как показатель устойчивости и неустойчивости спектральных характеристик речевого сигнала. В окрестностях локальных минимумов сигнал имеет устойчивые характеристики, что характерно для протяженных звуков, таких как гласные,

шипящие. В окрестностях локальных максимумов происходят изменения в структуре спектра, что характерно для границ между звуками и для взрывных звуков. К неустойчивым сегментам, проанализировав график на рис. 2, можно также отнести и хвостовые части протяженных звуков, что рассматриваемым алгоритмом принимается как переход к следующему сегменту. Исходя из вышесказанного, высотой максимумов в функции ошибки можно характеризовать степень неустойчивости сигнала в данный момент времени. Тогда, в терминах теории нечетких множеств, максимумы в функции ошибки восстановления можно интерпретировать как степень не принадлежности соответствующих векторов двум соседним классам и использовать эти сведения в алгоритмах кластеризации речевого сигнала как дополнительную информацию для уменьшения количества ошибок кластеризации.

Литература

1. Gustafsson F. Adaptive Filtering and Change Detection. Cloth, Wiley, 2001.
2. Ахмед Н., Рао К. Р. Ортогональные преобразования при обработке цифровых сигналов. – М.: Связь, 1980.
3. Патрик Э. Основы теории распознавания образов. – М.: Сов. радио, 1980.
4. Oja E. Principal components, minor components and linear neural networks // Neural Networks, 1992. – Vol 5. – Pp.927 – 935.

A SPEECH SIGNAL SPECTROGRAMS SEGMENTATION METHOD

Litvinenko S.

Volgograd State Technical University

The speech stream segmentation task is the important task in the speech recognition theory. It is solved a many different methods (see [1]). In this work we submit a method based on the speech signal analysis as an analysis of its power-density spectrum regarded as set of feature vectors which can be classified. But the final classification of segments is not performed and discrepancy function is calculated which is possessed the high value when the segment is span the feature vectors of two different classes. Here we use the assumption involves within the bounds of one class (a speech sound in the elementary case) the signal characteristics are being constant. Therefore a whole sound can be characterized by the single feature vector and its power modification function. A set of attribute space vectors is located in a single analysis window and relate to a single class will be locate nearly in the form of cloud which will be elongated more only in one direction.

The principal component analysis is applied to define this maximum variation direction. It constitute a linear orthogonal transform $Y = W \cdot X$ (see [2] and [3]), which is performed a data vector depended coordinate system rotation to arrange the axis along the vectors of $\boldsymbol{W}$ named the principal vectors. The first of this axis will be arranged along the maximum data variation. Therefore the first principal component $\boldsymbol{Y_1}$ will be characterized the signal power modification and the first principal eigenvector $\boldsymbol{W_1}$ will be corresponding to specific vector of given class. But if we try to describe two classes (two adjacent sounds) the first principal component axis will be arranged with the line connecting centers of two classes. In general case the first principal component axis will not characterize both the specific vector of first and second classes. Therefore the reconstruction of two adjacent sounds spectrogram only by first principal component will not be possible and we will have an error that will be more greatly then the error of reconstruction only single sound only by first principal component. Consequently the maximums of discrepancy function will be corresponding to sound transition regions.

Only the first principal component is finding in this method realization, for that one linear neuron is used. Its weights are learning on source data with using of expression suggested in [4]:

$$W_1(t+1) = W_1(t) + sY_1(X_i(t) - W_1(t)Y_1(t)) .$$

The spectrogram is examines by temporal sliding window for the discrepancy function definition. Then decomposition of spectrogram in this temporal window is performed by the neuron described above.

Next the restoration error is calculated after signal reconstruction. This process is performed for all temporal positions of this window and following this the reconstruction discrepancy function is received. This function has a local minimums then the window lies within the bounds of single sound and maximums then the window lies on the transition of two sounds.

References

5. Gustafsson F. Adaptive Filtering and Change Detection. Cloth, Wiley, 2001.
6. Ахмед Н., Рао К. Р. Ортогональные преобразования при обработке цифровых сигналов. – М.: Связь, 1980.
7. Патрик Э. Основы теории распознавания образов. – М.: Сов. радио, 1980.
8. Oja E. Principal components, minor components and linear neural networks // Neural Networks, 1992. – Vol 5. – Pp.927 – 935.