УДК 535.45; 535.785

НЕКОТОРЫЕ ОСОБЕННОСТИ РЕЧЕВЫХ СИГНАЛОВ

В.В. Митенок
Полесский государственный университет

Несмотря на многочисленные усилия на протяжении многих десятилетий, задача уверенного распознавания компьютером человеческой речи до сих пор не решена. Не решена также задача распознавания компьютером человека по его голосу, хотя известно, что хорошо знакомые между собой люди легко узнают один другого при разговоре по телефону. Более того, в последнее время прогресс в решении данной проблемы явно замедлился, причем существенно. Скорее всего, это связано с тем, что все те идеи, которые "лежали на поверхности", уже выработаны, полностью себя исчерпали, и для дальнейшего рывка вперед необходимы новые.

В большинстве случаев распознавание речи компьютером производится на основе разложения сигнала в спектр методом преобразований Фурье. Считается, что соотношение амплитуд разных мод определяет, какой именно звук звучит. Однако этот метод обладает рядом принципиальных недостатков. Отметим наиболее заметные из них, имеющие место не только при распознавании речи, но и при анализе сигналов любого происхождения:

1. Зависимость результатов от длительности сигнала
2. Уширение линий спектра
3. Наличие фиктивных (фальшивых) линий в спектре
4. Неоправданно усложненный вид спектра при дрейфе параметров.
5. Неустойчивость в отношении шумов, помех, погрешностей
6. Неуверенная идентификация слабых мод.

В противовес методу преобразований Фурье, в [1-2] был предложен метод аппроксимации для разложения сложного звука (звукового сигнала) на отдельные составляющие его слагаемые. Метод был задуман с целью учета дрожаний и дрейфа параметров звучания человеческого голоса, неизбежно присутствующих в речи, хотя бы в силу той причины, что реальный человеческий голос – неидеален. Даже оперные певцы способны выдерживать взятую ноту лишь с какой-то точностью. Поэтому игнорировать непостоянство параметров звучания никак нельзя. С другой стороны дрожания – это, скорее всего то, что, собственно, и позволяет узнавать человека по голосу. Если В предложенном методе аппроксимации заранее допускается, что звук представляет собой сумму мод с постоянными несущими частотами, но с медленно (по сравнению с несущими частотами) меняющимися амплитудами. Что касается дрожаний и дрейфа несущих частот и фаз, то их с помощью тривиальных математических преобразований можно перевести на непостоянство амплитуд.

Метод аппроксимации приводит к составлению систем линейных алгебраических уравнений относительно дрейфующих амплитуд и нуля отсчета, методы решений которых хорошо известны.

Для применения метода аппроксимации необходимо знать значения несущих частот. Эту проблему можно решить следующим образом. Сначала наугад выбираются некоторые числа в качестве значений частот, решается задача аппроксимации и подсчитывается остаточная невязка. Затем выбирается другой набор несущих частот, и также подсчитывается остаточная невязка. Затем остаточные невязки сравниваются, и, в качестве более правильного принимается тот набор частот, который обеспечивает меньшую остаточную невязку.

Поступая так многократно, можно выйти на тот набор (истинных) частот, который обеспечивает наименьшую невязку по сравнению с любым другим набором частот (или, по крайней мере, достаточно близко подойти к нему). Для ускорения процесса можно организовать перебор частот не случайным образом а, например, методом скорейшего спуска. Если же после нахождения истинных частот все же окажется, что остаточная невязка непомерно велика, то следует увеличить количество мод, и все расчеты повторить заново.

Разработанная методика применялась для изучения реальных звуков. Было проведено изучение образцов гласных звуков А, О, У, Ы, Э, а также долгие согласные Л,М,Н. полученных от 4-х мужчин и 4-х женщин (от 1000 до 5000 образцов для каждого звука от каждого респондента). Для более естественного звучания звуки произносились не сами по себе, а в виде мультислогов, каждый из которых состоял из 4-8 повторений одного и того же слога, составленного из одного из гласных и одного из согласных звуков из числа указанных выше.

При изучении образцов оказалось, что то значение базовой частоты, которое наблюдается из преобразований Фурье, должно быть уменьшено вдвое. Половинные частоты, как правило, обладают намного меньшей интенсивностью, нежели целые частоты, но они – есть. В большинстве случаев они проявляются в виде отдельных всплесков, далеко (на расстоянии порядка одного базового периода) отстоящие друг от друга. На спектрограмме Фурье они не видны в силу малой средней интенсивности, так как проигрывают фальшивым экстремумам. Вместе с тем в ряде случаев, приблизительно в одном из 70-80 образцов, половинные частоты имеют интенсивность большую, нежели целые. Но, так как это наблюдается достаточно редко, это, скорее всего, также остается незамеченным. Надо полагать, что двукратная ошибка в выборе базовой частоты – одна из причин замедления прогресса в задаче распознавания речи.

Широко распространено мнение, что человек, распознавая голос, речь, не реагирует на фазы мод. Наши исследования показали обратное - если звук представить в виде суммы мод с медленно меняющимися амплитудами

$$y_i = \sum_{k=1}^{l} A_{ki} \sin(\omega_k i + \varphi_k) \ , \qquad (1)$$

где A_{ki} – дрейфующие амплитуды мод, φ_k – их фазы, l – количество мод, (каждое слагаемое в (1) соответствует одной моде), и если из всех мод выделить две группы – так, чтобы суммы частот мод каждой из групп были одинаковы, и затем составить комбинацию

$$X = \sum_{i=1}^{n} \varphi_i^1 - \sum_{i=1}^{m} \varphi_i^2 \qquad (2)$$

из фаз отдельных мод, где φ^1 - фазы мод первой группы, φ^2 - фазы мод второй группы, n и m – соответственно количество мод в первой и во второй группе, (одни и те же фазы могут входить в (2) по несколько раз), то указанная комбинация, после разбраковки по базовым частотам, имеет намного меньшее среднее квадратичное отклонение, нежели при случайном наборе фаз.

Так, комбинация $X = 2\varphi_1 - \varphi_2$, усредненная по 4000 образцов, полученных от одного из респондентов показывает среднее отклонение, равное 0.23, а комбинация $X = \varphi_1 + \varphi_2 - \varphi_3$ - равное 0.39. Этим же свойством обладают и многие другие комбинации вида (2). Отметим, что если бы комбинация вида (2) имела равномерное распределение на интервале $[0,2\pi]$, то есть была бы случайной, то ее среднее квадратичное отклонение составляло бы $\frac{2\pi}{\sqrt{12}} \approx 1.81$. При переходе от звука к звуку среднее значение комбинации (2) меняется весьма заметно – во многих случаях до 1.2 – а их среднее отклонение остается практически тем же. Такая тесная группировка фазовых комбинаций вида (2) вокруг их средних значений при достаточно большом числе испытаний явно

не может быть случайной – более того, на основе анализа фаз нами была разработана компьютерная программа распознавания вышеупомянутых звуков, которая оказала уровень надежности в 90-95

Итак, при распознавании при распознавании звуков, фазы мод важны не сами по себе, а в виде комбинации (2). Это можно объяснить следующим образом: данная комбинация малочувствительна по отношению к выбору начала отсчета времени и к небольшим (до 10-15 процентов) ошибкам в выборе базовой частоты.

Литература:

1.В.В.Митянок "О числовых характеристиках некоторых низкочастотных звуков человеческой речи". Электронный журнал "Техническая акустика" www.ejta.org, N15, 2008

2. В.В.Митянок "Определение числовых характеристик высокочастотных звуков речи на основе аппроксимации гармоническими функциями". Известия НАН Беларуси. Сер.ф.-м.н. N2,2009, стр 111-118.