УДК 534.87; 534.4

О ГЕНЕРИРОВАНИИ ИСКУССТВЕННЫХ СЛОВ РЕЧИ ЧЕЛОВЕКА

Митянок Вячеслав Владимирович, к.ф.-м.н., доцент
Полесский государственный университет
Mitsianok Viachaslau Vladimirovich, PHD, mitsianok@mail.ru
Polessky State University

Проводились математические эксперименты по разложению звуков речи на моды и обратному их суммированию с целью выявления факторов, как значащих, так и незначащих для распознавания речи.

Ключевые слова: *распознавание речи, синтез речи, фазовый анализ звуков.*

Введение. Как хорошо известно, метод преобразований Фурье, используемый для анализа (квази)периодических сигналов, обладает рядом существенных недостатков [1,2]. Так, спектры сигналов являются размытыми, причем степень размытости зависит от длительности сигнала – при слишком малой длительности сигнала размытость линий становится настолько большой, что соседние линии спектра могут поглощать друг друга. При слишком большой длительности сигнала и при некоторой неустойчивости его параметров в спектре появляется множество фальшивых линий. Все это приводит к трудностям в задачах автоматического распознавания речи человека и верификации и идентификации личности по голосу. Косвенным признаком того, что метод преобразований Фурье не годится для этих задач является тот факт, что несмотря на значительные усилия и вложения денежных средств является то, что эти задачи до сих пор не имеют удовлетворительного решения. Достигнутые к настоящнму времени успехи можно считать лишь частичными.

В связи с этим в [3-5] был предложен метод аппроксимации, который предназначен для решения тех же задач, но который не имеет присущих методу преобразований Фурье недостатков. На основе метода аппроксимации был получен ряд принципиальных результатов. Так оказалось, что в спектре отдельных, долго произносимых звуков, присутствуют полуцелые (по отношению к базовой), частоты, действующие “вспышками”, имеет место “жесткая” модуляция амплитуд высших мод базовой частотой, причем со срывами. Тем самым было найдено объяснение неудачам метода преобразований Фурье.

В связи с имеющими место успехами метода аппроксимации имеет смысл применить его для создания искусственных звуков и слов речи человека. Если искусственные слова и звуки будут созданы, то тогда станет ясным, на что же именно следует обращать внимание при автоматическом распознавании речи, какие особенности звуковых сигналов позволяют отличать одного диктора от другого, а какие – наоборот, не имеют никакого значения, они случайны, привнесены несовершенством аппарата речеобразования человека, они лишь “путаются под ногами”, отвлекая внимание исследователей и заставляя их распылять свои усилия.

Метод аппроксимации. Метод аппроксимации основан на функционале [3-5]:

$$S = \sum_{i=1}^{n}\left[y(t_i) - y_1(t_i)\right]^2 + \alpha\sum_{k=1}^{n-1}(b_{0,i} - b_{0,i+1})^2 + \alpha\sum_{k=1}^{l}\sum_{i=1}^{n-1}(a_{k,i} - a_{k,ii+1})^2 + \alpha\sum_{k=1}^{l}\sum_{i=1}^{n-1}(b_{k,i} - b_{k,i+1})^2, \quad (1)$$

где $y(t_i)$ — зависящая от времени аппроксимируемая функция, описывающая сигнал, заданная своими значениями в n последовательных моментах времени от t_1 до t_n ,*a*

$$y_1(t_i) = b_{0,i} + \sum_{k=1}^{l}\left[a_{k,i}\sin(\omega_k t_i) + b_{k,i}\cos(\omega_k t_i)\right], \qquad i=1..n \quad (2)$$

— аппроксимирующая функция, $b_{0,i}$ — дрейфующий нуль (начало отсчета), $a_{k,i}$, $b_{k,i}$ — дрейфующие амплитуды синус и косинус волн (параметры аппроксимирующей функции), ω_k - их несущие частоты, l — количество волн (мод) в аппроксимирующей функции. Функционал (1) сконструирован как сумма слагаемых двух видов: слагаемые, не содержащие параметр α, отвечают за близость между аппроксимируемой и аппроксимирующей функциями, слагаемые, содержащие α, отвечают за сглаживание прыжков дрейфующих амплитуд волн (мод) аппроксимирующей функции при переходе по оси времени между соседними моментами дискретизации. Чем бóльшим выбрано значение α, тем более гладкими будут получаться амплитуды волн. Вычисляя частные производные (1) по дрейфующим амплитудам и по дрейфующему началу отсчета и приравнивая их нулю, получим систему линейных алгебраических уравнений относительно параметров аппроксимирующей функции. Решив эту систему, найдем эти параметры и тем самым произведем разложение аппроксимируемой функции на сумму волн с медленно меняющимися амплитудами. Найденные таким путем $b_{0,i}$, $a_{k,i}$, $b_{k,i}$ можно подставить в (2). Полученную в результате аппроксимирующую функцию можно назвать восстановленным звуком. Если затем вычесть восстановленный звук из исходного звука и подвергнуть разность преобразованиям Фурье, то часто выясняется, что существуют еще какие-то несущие частоты, которые не были замечены при первом разложении в ряд (интеграл) Фурье по причине малой интенсивности несомых ими мод. В частности, этим способом в [3-5] было установлено, что в спектре многих звуков присутствуют полуцелые (по отношению к базовой) несущие частоты. Соответственно, в звуке присутствуют целые и полуцелые, правда, малоинтенсивные, моды.

Каждую из мод, входящих в (2), можно переписать в физически более информативном виде:

$$a_{k,i}\sin(\omega_k t_i) + b_{k,i}\cos(\omega_k t_i) = c_{k,i}\sin(\omega_k t_i + \varphi_{k,i}), \qquad k=1..l, i=1..n \tag{3}$$

и тогда аппроксимирующая функция выглядит так:

$$y_1(t_i) = b_{0,i} + \sum_{k=1}^{l} c_{k,i}\sin(\omega_k t_i + \varphi_{k,i}). \quad i=1..n \tag{4}$$

Здесь $c_{k,i}$ - дрейфующая *общая* амплитуда волны (моды), $\varphi_{k,i}$ - дрейфующая фаза.

Синтез искусственных монозвуков. Исследовались те гласные звуки, которые можно было произносить долго – это звуки (монозвуки) «А», «О» , «У» , «Э» , «Ы» , «И» , полученные от нескольких респондентов, женщин и мужчин. Звуки раскладывались на моды и затем восстанавливались. Во всех случаях восстановленный звук звучал также как и исходный.

Для того, чтобы ответить на вопрос, *что* именно делает звук «А» звуком «А», звук «О» звуком «О» и т.д., перед суммированием (4) были проведены математические эксперименты по сознательному искажению амплитуд и (или) фаз. Так, фазы всех мод, кроме базовой, заменялись на искусственые, связанные с фазой базовой моды соотношением

$$\varphi_{k,i} = k\varphi_{1,i} + r_k, \qquad k=1..l,\ i=1..n \tag{5}$$

где k – номер моды, $\varphi_{1,i}$ - зависящая от времени фаза базовой моды, r_k – массив произвольных чисел. Фаза базовой моды не менялась. Не менялись и все дрейфующие амплитуды. Как оказалось, звучание звуков от такой замены не изменилось.

Выяснилось также, что при суммировании можно опустить дрейфующий нуль и полуцелые моды. И от такого отбрасывания звук не менялся. А вот если фазу каждой из мод, в том числе и базовой, на всем отрезке звучания заменить на постоянное, но случайное число, то качество звука значительно ухудшалось. Вместо четкого звука слышалось то, что скорее можно назвать звучанием зуммера.

В поисках объяснений этому явлению были проделаны следующие математические эксперименты. Усредненный амплитудный спектр каждого из изучаемых звуков соединялся в формуле (4) с дрейфующими фазами от любого другого из этих же звуков и от любого из других респондентов. После такой операции звук не менялся, звучал четко и соответствовал именно амплитудам. Как это можно объяснить? Оказалось, что во всех случаях реальные фазы не являются стро-

гими константами, а дрейфуют (плавают) вокруг неких средних значений с неустойчивым периодом от 1.5 до 2.5 Гц и с неустойчивой амплитудой 0.5-2 радиан. Они – как бы «испорчены». В связи с этим возникло предположение, что именно так и должно быть. Что мозг слушателя уже готов к тому, что диктор будет производить сигнал с испорченной фазой, а звук с неиспорченной фазой мозгом слушателя за звук не воспринимается. Это предположение оправдалось. Когда в качестве фазы принималась хаотически меняющаяся (в определенных рамках) величина, то звук вновь звучал четко и распознаваемо.

Таким образом, для синтеза вышеуказанных звуков, вместо (4), как один из вариантов, можно принять выражение

$$y_1(i) = \sum_{k=1}^{l} c_k \sin(\omega_k i + kp \sin(i/3300) + r_i \quad i=1..n \qquad (6)$$

где ω_k– несущие частоты, пропорциональные базовой, r_j – массив произвольных чисел, n – длина отрезка звучания (в отсчетах дискретизации). Множитель p в (6) может принимать любое значение в интервале [1..10], но лучшее звучание наблюдается при p=2 для звуков «Э», «Ы», и p=4 для звуков «А», «О», «У», «И». За основу получения усредненных общих был взят голос автора. Внутренний синус в (6) обеспечивает порчу фазы. (Возможны и другие варианты порчи фаз.)

Список использованных источников

1.Васильева Л. Г., Жилейкин Я. М., Осипик Ю. И. Преобразования Фурье и вейвлет-преобразования. Их свойства и применение. //Вычислительные методы и программирование: в 3 т. – М., – Т 3, – Вып 1, – С 172-175, 2002

2. Митянок В. В. Определение числовых характеристик высокочастотных звуков речи на основе аппроксимации гармоническими функциями // Известия НАН Беларуси, сер ф.-м.н., – Минск, – № 2. –С 111-118. 2009

3. Митянок В. В., Коновалова Н. В. Применение фазового анализа звуков речи для распознавания человека по его голосу. [Электронный ресурс] //Техническая акустика. – Электрон. журн.- СПб., – 2013. № 4. – Режим доступа: http://www.ejta.org, свободный

4.Митянок В. В. О числовых характеристиках некоторых низкочастотных звуков человеческой речи. [Электронный ресурс] // Техническая акустика. – Электрон. журн. – СПб., 2008. – № 15. – Режим доступа: http://www.ejta.org, свободный

5. Митянок В. В. К проблеме идентификации и верификации личности по фазовым характеристикам звуков речи [Электронный ресурс] //Техническая акустика. – Электрон. журн. – СПб., – 2015. – № 7. – Режим доступа: http://www.ejta.org, свободный