Синтез речи

Си́нтез ре́чи — в широком смысле — восстановление формы речевого сигнала по его параметрам; в узком смысле — формирование речевого сигнала по печатному тексту. Часть искусственного интеллекта.

Синтезом речи — прежде всего называется всё, что связано с искусственным производством человеческой речи.

Синтезатор речи — структура, способная переводить текст/образы в речь, в программном обеспечении или аппаратных средствах.

Голосовой движок — непосредственно система/ядро преобразования текста/команд в речь, это также может существовать независимо от компьютера.

Способы синтеза речи

Все способы синтеза речи можно подразделить на группы

- параметрический синтез;
- компиляционный (компилятивный) синтез;
- синтез по правилам;

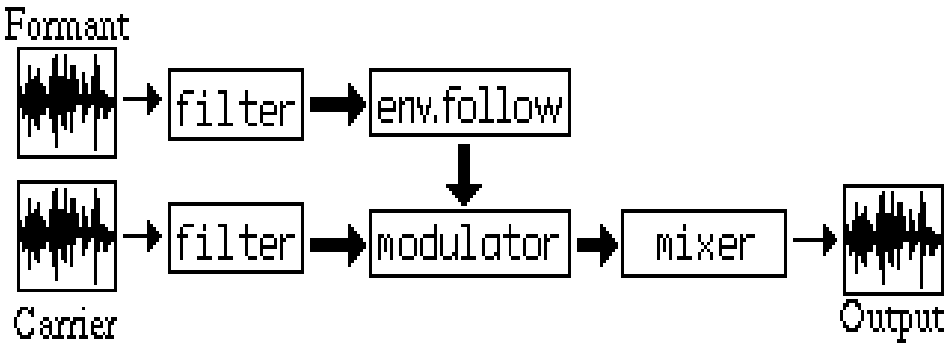
Параметрический синтез

Параметрический синтез речи является конечной операцией в вокодерных системах, где речевой сигнал представляется набором небольшого числа непрерывно изменяющихся параметров. Достоинством такого способа является возможность записать речь для любого языка и любого диктора. Качество параметрического синтеза может быть очень высоким (в зависимости от степени сжатия информации в параметрическом представлении).

Вокодер (*voice coder* — кодировщик голоса) — устройство синтеза речи на основе произвольного сигнала с богатым спектром. Изначально вокодеры были разработаны в целях экономии частотных ресурсов радиолинии системы связи при передаче речевых сообщений. Экономия достигается за счёт того, что вместо собственно речевого сигнала передают только значения его определённых параметров, которые на приёмной стороне управляют синтезатором речи. Основу синтезатора речи составляют три элемента:

- генератор тонального сигнала для формирования гласных звуков;
- генератор шума для формирования согласных;

- и система формантных фильтров для воссоздания индивидуальных особенностей голоса.

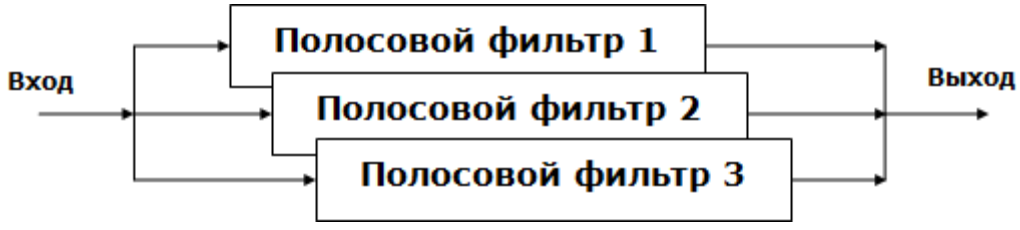


Формантный фильтр является одним из ключевых компонентов в системах синтеза речи и речеподобных сигналов.

Формантный фильтр — система резонансных фильтров, предназначенная для генерации речевого сигнала с заданной фонетической структурой.

В основу структуры формантного фильтра заложена упрощённая модель голосового тракта. В соответствии с моделью, голосовой тракт представляет собой резонатор с несколькими пиками АЧХ, частоты которых определяют вид произносимой фонемы. Эти пики АЧХ получили название форманты. Пример спектра фонемы «А»:

Формантный фильтр создаёт формантные области в спектре входного сигнала с помощью нескольких параллельно соединённых полосовых фильтров. Количество звеньев в схеме определяет порядок формантного фильтра. Схема формантного фильтра третьего порядка:



Чтобы синтезировать речевой сигнал, соответствующий определённой фонеме, необходимо настроить центральную частоту каждого полосового фильтра системы на соответствующую частоту форманты. Таблица частот формант для некоторых фонем (приведённые значения для других голосов, как правило, отличаются):

Фонема	**Первая форманта, Гц**	**Вторая форманта, Гц**	**Третья форманта, Гц**
«и»	270	2300	3000
«е»	400	2000	2550
«а»	660	1700	2400
«у»	640	1200	2400

Входным воздействием для формантного фильтра могут служить различные сигналы с различной окраской тембра. Основными параметрами входного сигнала являются частота повторения и ширина спектра. Частота повторения определяет высоту тона синтезируемой фонемы и лежит в пределах от 200 Гц до 2000 Гц. При этом нижние частоты данного диапазона соответствуют тембру мужского голоса, верхние — женского. Занимаемый входным сигналом диапазон частот должен быть как можно шире. В качестве такого сигнала часто используется импульсная последовательность.

После всех преобразований голос человека становится похожим на "голос робота", что вполне терпимо для средств связи и интересно для музыкальной сферы. Так было лишь в самых примитивных вокодерах 1-й половины XX ст. Современные связные вокодеры обеспечивают высочайшее качество голоса при существенно более сильной степени сжатия в сравнении с упомянутыми выше.

Вокодер как необычный эффект был взят на вооружение электронными музыкантами и впоследствии стал полноценным эффектом благодаря фирмам-изготовителям музыкального оборудования, которые придали ему форму и удобство музыкального эффекта. Вокодер как музыкальный эффект позволяет перенести свойства одного (модулирующего) сигнала на другой сигнал, который называют носителем. В качестве сигнала-модулятора используется голос человека, а в качестве носителя — сигнал, формируемый музыкальным синтезатором или другим музыкальным инструментом. Так достигается эффект «говорящего» или «поющего» музыкального инструмента. Помимо голоса, модулирующий сигнал может быть и гитарой, клавишными, барабанами и вообще любым звуком синтетического и «живого» происхождения. Также нет ограничений и на несущий сигнал. Экспериментируя с моделирующим и несущим сигналом, можно получать

совершенно разные эффекты — говорящая гитара, барабаны со звуком фортепиано, гитара, звучащая как ксилофон.

Современные вокодеры можно поделить на:

1. Аппаратные (с которых всё и началось)
2. Виртуальные (появились гораздо позже с развитием компьютерных технологий создания музыки).

В практике компьютерного музыканта значительно чаще используются вокодеры виртуальные, реализованные в виде VST-плагинов (VST-plugins),

Virtual Studio Technology (**VST**) — формат ресурсозависимых (*native*) плагинов реального времени, которые подключаются к звуковым редакторам и музыкальным редакторам, секвенсорам и т. д.

Применяются подобные решения вокодеров как самостоятельные программы, так и совместно с программой-хостом. В качестве таковой может быть использована любая виртуальная студия, поддерживающая технологию VST, например, Sound Forge, Steinberg Cubase или FL Studio. Программа-хост позволяет подключать собственно сам вокодер и выбирать, откуда будет поступать несущий и модулирующий сигнал (некоторые вокодеры имеют встроенный синтезатор несущего сигнала) — с синтезаторов и семплеров (которые, кстати, имеют тоже формат VST) или с микрофонов и других подключённых к звуковой карте инструментов. А управление несущим сигналом осуществляется посредством MIDI-команд, поступающих из MIDI-секвенсора или MIDI-клавиатуры в VST-плагин (синтезатор или семплер).

Примером виртуальных вокодеров могут служить VST-плагины, такие как: Steinberg Vocoder, Fruity Vocoder, Akai DC Vocoder, Voctopus, AC vocoder, Formulator, Lpc-vocoder, Darkoder, Cylonix (как работающий самостоятельно (standalone)). Любой знакомый с его принципиальным устройством может собрать собственный вокодер в любой модульной программе типа NI Reactor/Generator, Max MSP, Buzz Composer.

Компиляционный синтез

Компиляционный синтез сводится к составлению сообщения из предварительно записанного словаря исходных элементов синтеза. Размер элементов синтеза не меньше слова. Очевидно, что содержание синтезируемых сообщений фиксируется объёмом словаря. Основная проблема в компилятивном синтезе — объёмы памяти для хранения словаря. В связи с этим используются разнообразные методы сжатия/кодирования речевого сигнала. Компилятивный синтез имеет широкое практическое применение.

Системами речевого ответа оснащаются разнообразные устройства (от военных самолётов до бытовых устройств) сейчас они находят всё большее применение в повседневной жизни, например, в справочных службах операторов сотовой связи при получении информации о состоянии счета абонента.

Полный синтез речи по правилам

Полный синтез речи по правилам (или синтез по печатному тексту) обеспечивает управление всеми параметрами речевого сигнала и, таким образом, может генерировать речь по заранее неизвестному тексту. В этом случае параметры, полученные при анализе речевого сигнала, сохраняются в памяти так же, как и правила соединения звуков в слова и фразы. Синтез реализуется путём моделирования речевого тракта, применения аналоговой или цифровой техники. Причём в процессе синтезирования значения параметров и правила соединения фонем вводят последовательно через определённый временной интервал, например 5—10 мс. Метод синтеза речи по печатному тексту (синтез по правилам) базируется на запрограммированном знании акустических и лингвистических ограничений и не использует непосредственно элементы человеческой речи. В системах, основанных на этом способе синтеза, выделяется два подхода. Первый подход направлен на построение модели речепроизводящей системы человека, он известен под названием *артикуляторного синтеза*. Второй подход — *формантный синтез по правилам*. Разборчивость и натуральность таких синтезаторов может быть доведена до величин, сравнимых с характеристиками естественной речи.

Синтез речи по правилам с использованием предварительно запомненных отрезков естественного языка — это разновидность синтеза речи по правилам, которая получила распространение в связи с появлением возможностей манипулирования речевым сигналом в оцифрованной форме. В зависимости от размера исходных элементов синтеза выделяются следующие виды синтеза:

- микросегментный (микроволновый);
- аллофонический;
- дифонный;
- полуслоговой;
- слоговой;
- синтез из единиц произвольного размера.

Обычно в качестве таких элементов используются полуслоги — сегменты, содержащие половину согласного и половину примыкающего к нему гласного. При этом можно синтезировать речь по заранее не заданному тексту, но трудно управлять интонационными характеристиками. Качество такого синтеза не соответствует качеству естественной речи, поскольку на границах

сшивки дифонов часто возникают искажения. Компиляция речи из заранее записанных словоформ также не решает проблемы высококачественного синтеза произвольных сообщений, поскольку акустические и просодические (длительность и интонация) характеристики слов изменяются в зависимости от типа фразы и места слова во фразе. Это положение не меняется даже при использовании больших объёмов памяти для хранения словоформ.