Министерство образования Российской Федерации
Владимирский государственный университет
Кафедра радиотехники и радиосистем

ЭФФЕКТИВНОЕ КОДИРОВАНИЕ И РАСПОЗНАВАНИЕ РЕЧЕВЫХ СИГНАЛОВ

Методические указания к лабораторным работам

Составитель
Е.К. ЛЕВИН

Владимир 2002

УДК 621.395

Рецензент
Кандидат технических наук, доцент
Владимирского государственного университета
А.Д. Поздняков

Печатается по решению редакционно-издательского совета
Владимирского государственного университета

Эффективное кодирование и распознавание речевых сигналов: Метод. указания к лабораторным работам / Владим.гос.ун-т: Сост. Е.К. Левин. Владимир, 2002. 51 с.

Рассматриваются вопросы экспериментального исследования цифровой обработки речевых сигналов, которая используется в цифровой телефонии при кодировании и автоматическом распознавании сигналов. Исследования проводятся на программных моделях соответствующих устройств. С помощью моделей определяются оптимальные параметры речевых кодеков и устройств автоматического распознавания голосовых команд.

Методические указания предназначены для студентов старших курсов направления «Радиотехника», специальностей 071500 – радиофизика и электроника, 200700 – радиотехника, 201500 – бытовая радиоэлектронная аппаратура дневной и заочной форм обучения.

Ил.13. Библиогр.: 4 назв.

УДК 621.395

Введение

Передача речевых сигналов по цифровым каналам связи подразумевает использование средств эффективного кодирования сигналов, которые обеспечивают уменьшение объема – сжатие потоков данных.

Разработка устройств эффективного кодирования - речевых кодеков - характеризуется большой трудоемкостью. Трудоемкость обусловлена, во-первых, сложностью алгоритмов, используемых для обработки сигналов, а во-вторых, большим объемом экспериментальных исследований разрабатываемых устройств. Последнее обстоятельство связано с субъективным характером восприятия искажений речевого сигнала, возникающих при его сжатии. С целью снижения трудоемкости разработки необходимо в ее процессе использовать специальные программно-аппаратные средства. Данные средства позволяют оценить искажения на слух и визуально (по «осциллограммам») на различных стадиях обработки сигналов в кодеке. Упрощается процесс отладки программного обеспечения кодека.

Необходимость использования специальных программно-аппаратных средств возникает и при разработке устройств распознавания речевых сигналов, применяемых, в частности, в компьютерной телефонии при организации удаленного доступа к данным с помощью голосовых команд. Программное обеспечение устройств реализует сложные алгоритмы определения параметров речевых сигналов, сопоставления их с эталонами голосовых команд, создания эталонов. Сильная изменчивость речевых сигналов обусловливает появление ошибок распознавания. Поэтому при разработке устройств много времени тратится на их экспериментальные исследования. Указанные средства позволяют ускорить процесс выявления причин ошибок распознавания, снизить трудоемкость создания эталонов (моделей) голосовых команд. Комплекс лабораторных работ позволяет студентам приобрести навыки практической работы с указанными средствами, а также дает возможность проанализировать сложные процессы обработки речевых сигналов.

1. Лабораторная работа

Исследование анализирующего фильтра речевого кодера на основе линейного предсказания

Цель работы

Исследование зависимости искажений сигнала, возникающих при его эффективном кодировании, от точности анализа – определения огибающей кратковременного спектра сегмента сигнала.

Краткие теоретические сведения

Работа современных речевых кодеков основана на следующей модели речеобразования. Речевой сигнал (РС) рассматривается как реакция фильтра-модели голосового тракта (синтезирующего фильтра) на сигнал возбуждения, который моделирует работу голосовых связок и процесс поступления воздушного потока в голосовой тракт. Данные модели характеризуются небольшим количеством параметров. Передача их по каналу связи вместо цифровых отсчетов сигнала значительно сокращает поток данных. Дополнительное сокращение потока достигается кодированием параметров РС – представлением их малоразрядными двоичными числами.

При кодировании РС определяются параметры огибающей кратковременного спектра сигнала – модели голосового тракта, а также параметры сигнала возбуждения.

Модели приближенно описывают процесс формирования РС, поэтому при декодировании появляются искажения сигнала. Свою долю искажений в сигнал вносит и кодер его параметров (за счет погрешности квантования параметров).

В лабораторной работе анализируются искажения сигнала, обусловленные лишь погрешностями моделирования голосового тракта – определения параметров огибающей его кратковременного спектра.

Огибающая кратковременного спектра сигнала представляется набором коэффициентов частной корреляции (КЧК), определяемых с помощью

анализирующего сигнал (в данном случае лестничного) фильтра. Работа фильтра базируется на идее линейного предсказания последующего отсчета сигнала на основе знания нескольких его предыдущих отсчетов. Чем больше используется предыдущих отсчетов - звеньев фильтра (коэффициентов частной корреляции), тем точнее определяется огибающая спектра и тем меньше искажения синтезированного сигнала (сигнала на выходе речевого кодека).

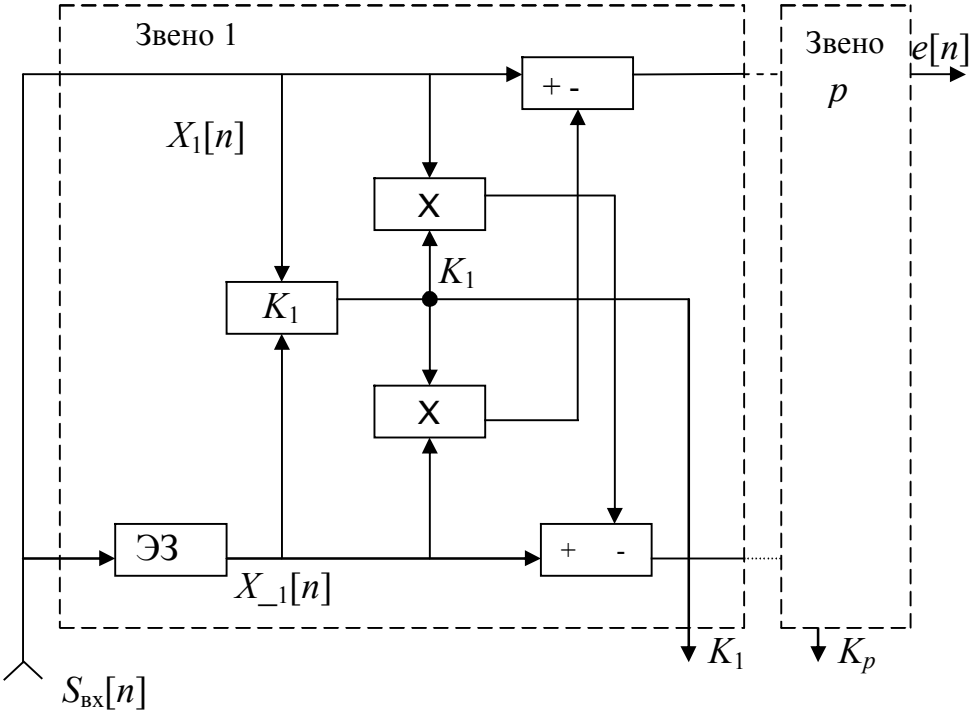


Рис.1.1

В данном случае исследуется изображенный на рис. 1.1 лестничный фильтр, который характеризуется набором (вектором) $\boldsymbol{K} = \{K_1, K_2 \ldots K_p\}$ коэффициентов частной корреляции:

$$K_i = \left(\sum_{n=1}^{N} X_i[n] X_{-i}[n] \right) \left(\sum_{n=1}^{N} X^2{}_i[n] \sum_{n=1}^{N} X^2{}_{-i}[n] \right)^{-0,5}, \; i \in \{1 \ldots p\}, \; (1.1)$$

где p – число звеньев фильтра; N – количество временных выборок сегмента; $X_i[n]$, $X_{-i}[n+1]$ – значения входных сигналов для i-го звена в моменты времени $nT_д$, $(n+1)T_д$; $T_д$ – период дискретизации (здесь –

125 мкс); $S_{вх}[n]$, $e[n]$ – входной и выходной сигналы анализирующего фильтра; ЭЗ – элемент задержки на величину $T_д$; Х – перемножитель сигналов.

Исследование фильтра целесообразно проводить, используя модель сегмента речевого сигнала, параметры которой можно быстро менять в широких пределах.

Структурная схема обработки сигналов при исследовании анализирующего фильтра приведена на рис. 1.2.

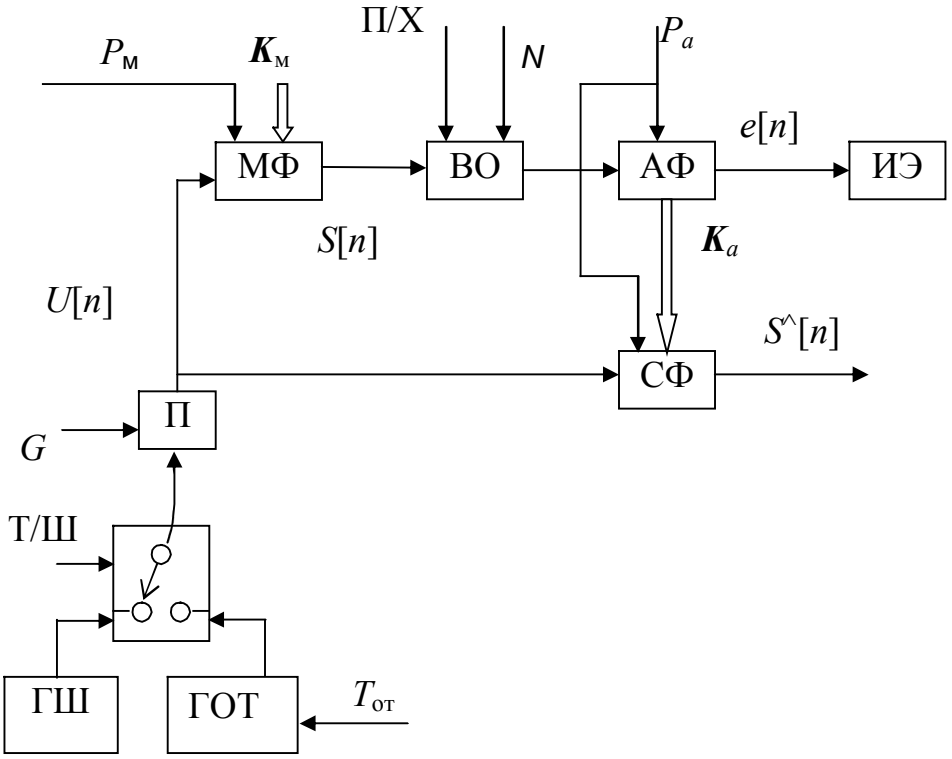


Рис.1.2

В соответствии с требуемым для исследования типом звука (невокализованный или вокализованный) для формирования модели сегмента используется либо генератор шума ГШ, либо генератор импульсов основного тона ГОТ с заданным периодом $T_{от}$ (или частотой $f_{от}$) следования импульсов.

Подключение нужного генератора осуществляется ключом по команде «Т/Ш». С выхода ключа сигнал поступает на перемножитель П, который

меняет уровень сигнала в соответствии с заданным коэффициентом усиления звука G.

Далее сигнал проходит через фильтр МФ с числом звеньев $P_{\text{м}}$, моделирующий голосовой тракт. С помощью вектора коэффициентов $\boldsymbol{K}_{\text{м}}$ фильтра формируется требуемая огибающая спектра сегмента модели речевого сигнала. Чем больше величины $P_{\text{м}}$ и P_a, тем точнее моделируется сегмент речевого сигнала и меньше средняя энергия E_e сигнала остатка предсказания на выходе АФ.

С выхода фильтра МФ сигнал $S[n]$ проходит через временное окно ВО, которое выделяет лишь определенный временной отрезок сигнала – сегмент.

Размер сегмента задается числом N анализируемых временных отсчетов сигнала. Часто используемые формы окна – прямоугольное и окно Хэмминга. Простое по реализации – прямоугольное окно, а окно Хэмминга обеспечивает более высокую точность оценки огибающей спектра сигнала по сравнению с прямоугольным. Установка типа окна осуществляется по команде «П/Х». После ВО сигнал попадает на анализирующий фильтр АФ (с количеством звеньев P_a) и далее - на измеритель средней энергии остатка $e[n]$ предсказания ИЭ.

Вычисляемый при работе анализирующего фильтра АФ набор из P_a коэффициентов частной корреляции $\boldsymbol{K_a}$ далее используется для формирования частотной характеристики синтезирующего фильтра СФ. Выход СФ представляет собой синтезированный – восстановленный по набору коэффициентов $\boldsymbol{K_a}$ – сигнал, использующий тот же сигнал возбуждения, что и при формировании модели сигнала. Если вектор $\boldsymbol{K_a}$ точно описывает огибающую спектра сигнала, то синтезированный сигнал $S^{\wedge}[n]$ мало отличается от исходного $S_{\text{вх}}[n]$.

Оценка огибающей кратковременного спектра производится с некоторыми погрешностями, которые обусловливают искажения синтезированного сигнала. Погрешность растет с уменьшением количества звеньев P_a , что подтверждается зависимостью средней энергии остатка предсказания на выходе АФ от P_a .

$$E_e(P_a) = E_S\,(1 - K_1^2)\,(1 - K_2^2)\ldots\,(1 - K_{Pa}^2), \qquad (1.2)$$

где E_S – средняя энергия сигнала на входе АФ; $K_1,\ldots K_{Pa}$ – коэффициенты частной корреляции в АФ. Чем больше количество КЧК, тем меньше энергия остатка предсказания и тем меньше погрешность.

Снижение погрешности путем чрезмерного увеличения P_a приводит к неоправданному росту объема вычислений. На практике обычно $P_a = 8\ldots12$.

Длительность ВО целесообразно установить равной небольшому количеству периодов наиболее низкочастотного основного тона (основной тон характеризует колебания голосовых связок). Чрезмерное уменьшение N приводит к значительным отличиям кратковременного спектра от долговременного в частотных точках кратных частоте основного тона, по которым определяется огибающая спектра. Чрезмерное увеличение N приводит к сильной изрезанности кратковременного спектра, что может вызвать увеличение погрешности определения сглаженной огибающей спектра с помощью небольшого числа коэффициентов частной корреляции. Кроме того, чрезмерное увеличение N приводит к нарушению условия квазистационарности РС на анализируемом сегменте сигнала.

Частотная характеристика окна Хэмминга характеризуется меньшим уровнем боковых лепестков по сравнению с прямоугольным окном. Поэтому кратковременный спектр сегмента сигнала, полученный с помощью окна Хэмминга, в меньшей степени отличается от долговременного спектра РС нежели в ситуации с прямоугольным окном. Этим обстоятельством и объясняется преимущественное использование окна Хэмминга на практике. Следует однако учитывать, что ширина главного лепестка частотной характеристики окна Хэмминга примерно в два раза больше лепестка частотной характеристики прямоугольного окна. Поэтому разрешающая способность по частоте у окна Хэмминга ниже.

Искажения сигнала, возникающие при его сжатии, можно оценить объективными и субъективными методами. Определяющей является субъективная оценка искажений на слух, так как именно она характеризует восприятие человеком синтезированных звуков речи. Используются также визуальные оценки искажений по осциллограммам и спектрограммам сигнала. Можно также сравнивать осциллограммы остатка предсказания и сигнала возбуждения в модели речевого сигнала. Чем более они похожи, тем меньше искажения. Однако субъективные оценки трудно поддаются

математическому описанию, поэтому их использование в алгоритмах кодирования затруднительно.

Объективная оценка искажений сигнала производится путем определения дисперсии остатка предсказания на выходе анализирующего фильтра. Чем меньше дисперсия, тем точнее фильтр определяет огибающую спектра сигнала.

Для объективной оценки можно также использовать сумму квадратов разностей между преобразованными значениями исходных значений КЧК и значениями КЧК, определенными при анализе. Эта оценка характеризует искажения огибающей кратковременного спектра сигнала.

Порядок выполнения работы

1. Провести анализ смоделированного сегмента речевого сигнала для заданного звука при различных значениях порядка P_a анализирующего фильтра (порядка предсказания). Построить графики зависимостей искажений $A(P_a)$ огибающей кратковременного спектра и нормированной средней энергии $E_e^{\text{н}}(P_a)$ остатка предсказания от порядка анализирующего фильтра.

$$E_e^{\text{н}}(P_a) = \frac{E_e(P_a)}{E_e(1)},$$

где $E_e(P_a)$ – средняя энергия остатка предсказания.

Используя выражение (1.2) и определенные при анализе значения КЧК, рассчитать зависимость нормированной средней энергии остатка предсказания от порядка предсказания.

$$E_{e\,p}^{\text{н}}(P_a) = (1-K_2^2)(1-K_3^2)...(1-K_{Pa}^2),$$

где K_i – КЧК для i-го звена лестничного фильтра. Построить график зависимости $E_{e\,p}^{\text{н}}(P_a)$ и сравнить его с графиком зависимости $E_e^{\text{н}}(P_a)$, определенной в процессе эксперимента.

2. Сравнить «осциллограммы» смоделированного для анализа сегмента сигнала и сегмента, синтезированного по КЧК, определенным при анализе для различных значений порядка P_a предсказания.

По результатам сравнения определить минимальное значение $P_a = P_{\min}$, при котором отличия сравниваемых сигналов являются незначительными

(экспертная оценка).

Зафиксировать «осциллограммы» для $P_a = P_{\min}$, а также для $P_a = 1$ (случай максимальных искажений сигнала).

При частоте основного тона 1Гц и $N = 511$ зафиксировать спектрограммы для $P_a = P_{\min}$ и $P_a = 1$, которые представляют огибающие кратковременных спектров синтезированного сигнала для различных значений порядка предсказания.

3. Сравнить «осциллограммы» остатка предсказания и сигнала возбуждения моделирующего фильтра для различных значений порядка предсказания. Зафиксировать «осциллограммы» для $P_a = P_{\min}$.

4. Для $P_a = 10$ построить график зависимости средней энергии остатка предсказания от размера временного окна $E_e(N)$.

5. Провести экспертную оценку искажений речевого сигнала, возникающих при его сжатии, на слух:

- создать или использовать готовый тестовый звуковой файл;
- «пропустить» его через речевой кодек для $P_a = P_{\min}$ и записать в файл декодированный сигнал;
- повторить предыдущий пункт работы для $P_a = 10$;
- прослушать созданные записи речевых сигналов и оценить изменение уровня искажений сигнала при изменении порядка предсказания, учитывая разборчивость речи и узнаваемость диктора по голосу.

6. Рассчитать коэффициент сжатия при $P_a = P_{\min}$ и $P_a = 10$.

7. Сделать выводы по работе.

Содержание отчета

1. Структурная схема процесса исследования анализирующего фильтра.
2. Значения исследуемых зависимостей, их графики.
3. «Осциллограммы».
4. Выводы по работе.

Контрольные вопросы

1. Почему нельзя устанавливать длительность временного окна чрезмерно большую?

2. Какова форма сигнала на выходе анализирующего фильтра высокого порядка в случае вокализованного звука?
3. Почему число звеньев анализирующего фильтра не может быть большим?
4. Как связана структура речевого кодека с моделью голосового тракта?
5. Почему нельзя устанавливать длительность временного окна чрезмерно малую?
6. Какова роль объективных и субъективных оценок искажений сигнала при проектировании речевых кодеков?
7. Почему при анализе речевого сигнала используется перекрытие временных окон?
8. В чем заключается отличие синхронного метода анализа от асинхронного?
9. Почему при анализе речевого сигнала обычно используется окно Хэмминга?
10. В чем заключается различие между результатами спектрального анализа, проведенного методом Фурье и на основе линейного предсказания?
11. Если окно Хэмминга и прямоугольное окно имеют равные длительности, то какое из них имеет большую разрешающую способность по частоте?

2. Лабораторная работа

Исследование измерителя основного тона речевого сигнала

Цель работы

Исследовать зависимость параметров кратковременной функции средней разности от отношения сигнал – шум, определить условие классификации сегментов сигнала на вокализованные и невокализованные.

Краткие теоретические сведения

Назначение измерителя основного тона ИОТ — классификация сегментов речевых сигналов на невокализованные и вокализованные, а также определение периода (частоты) основного тона в последнем случае.

Принцип работы ИОТ основан на анализе автокорреляционной функции сигнала или связанной с ней функции с целью определения их периода. В последнее время широко используется кратковременная функция среднего значения разности – *AMDF* (*Average Magnitude Difference Function*):

$$y(k)=\frac{1}{R}\sum_{n=0}^{N-1}\left|X[n]\ -X[n+k]\ \right|,\quad R=\sum_{n=0}^{N-1}\left|X[n]\right|,$$

где R – нормирующий делитель; $X[n]$ – значение входного сигнала ИОТ в момент времени $nT_д$; $T_д$ – период дискретизации; N – число выборок в сегменте сигнала. В общем случае $X[n]$ – сумма периодического и случайного компонентов, поэтому данная функция является случайной. Типичная форма ее математического ожидания для вокализованных звуков изображена на рис. 2.1 (сплошная линия). Штриховыми линиями указана зона наиболее вероятных значений $y(k)$.

Период $T_{от}$ основного тона определяется расстоянием между двумя минимумами функции. Измерительный порог $r_и$ служит для фиксации минимумов. Минимальные значения функции определяются лишь для тех значений k, которые обеспечивают выполнение условия $y(k) \le r_и$. Из

рис. 2.1. следует, что чем больше разброс значений $y(k)$, меньших $r_и$, тем больше разброс значений измеренного периода (T_{min} … T_{max}), а следовательно, и больше погрешность измерения. Если шумовая составляющая в сигнале достаточно велика, то принимается решение о невокализованности звука. При этом наименьшее значение функции превышает «классификационный» порог $r_к$. Необходимо отметить, что значения $y(k)$ в точках локальных минимумов определяются не только периодичностью исследуемого сигнала, то есть соотношением шумового и детерминированного компонентов в данном случае, но и уровнем формант.

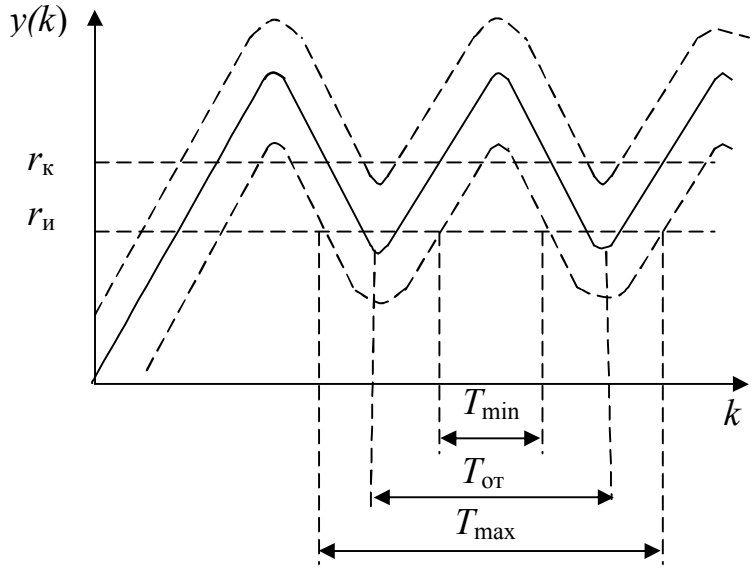


Рис. 2.1

Если форманты достаточно мощные, то появляются дополнительные глубокие «провалы» в функции, которые могут привести к грубым ошибкам измерения периода основного тона $T_{от}$. Поэтому часто на вход ИОТ подается не сам сегмент сигнала, а его остаток предсказания на выходе анализирующего фильтра. В остатке предсказания формантная структура сигнала значительно разрушена при сохранении периодичности, обусловленной импульсами основного тона.

Так как наиболее вероятные значения частоты основного тона лежат в пределах (50…500)Гц, то для повышения точности измерения перед

вычислением *AMDF* сигнал пропускают через фильтр нижних частот (ФНЧ) с частотой среза 1 кГц.

Схема обработки сигналов в лабораторной работе

На рис. 2.2 приведена схема обработки сигналов, проводимой в лабораторной работе. Фильтр нижних частот ФНЧ, вычислитель *AMDF* и измеритель И, где определяется признак вокализованности сегмента «Т/Ш» и временной интервал между соседними минимумами *AMDF*, составляют измеритель основного тона ИОТ. С помощью генераторов: шума – ГШ, импульсов – ГИ и гармонических колебаний – Г формируется модель сегмента сигнала на входе ИОТ с различным отношением сигнал - шум. Оно изменяется с помощью перемножителя П в зависимости от коэффициента усиления *G*. По команде «И/С» (импульс – синусоидальный сигнал) ключ подключает либо ГИ – при этом моделируется ситуация, когда на входе ИОТ присутствует остаток предсказания, либо Г – при этом моделируется случай анализа вокализованного сегмента непосредственно.

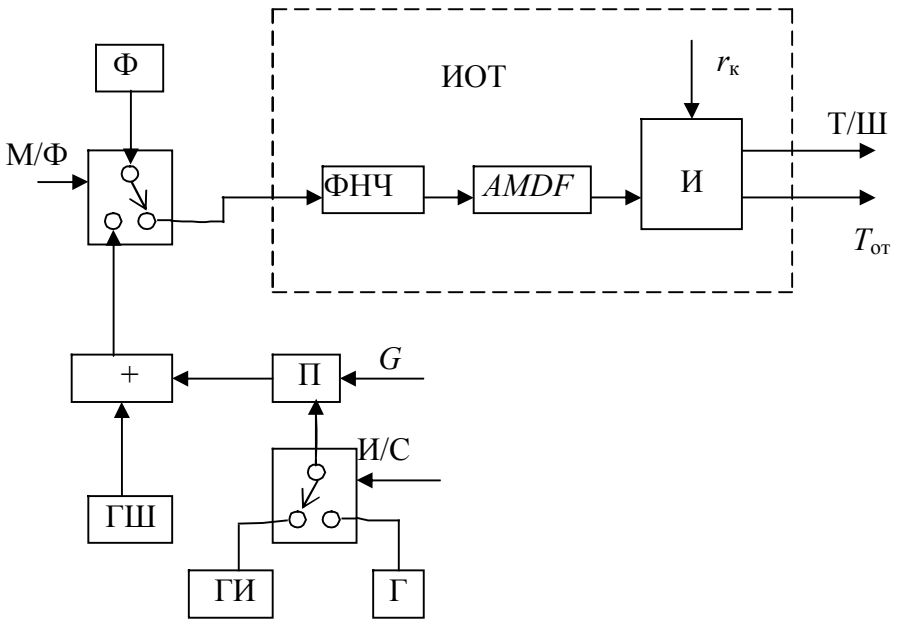


Рис. 2.2

В лабораторной работе имеется также возможность подачи на вход ИОТ сигнала из файла Ф записанного заранее реального звука. При этом можно оценить качество работы ИОТ с реальными сигналами.

Подключение файла Ф осуществляется по команде «М/Ф» (модель – запись реального сигнала). Фильтр нижних частот ФНЧ, вычислитель функции *AMDF* и измеритель периода основного тона И образуют собственно структуру ИОТ.

В лабораторной работе предусмотрена возможность наблюдения «осциллограмм» сигналов на входе ИОТ, а также графика *AMDF*.

Порядок выполнения работы

1. Для заданной формы входного сигнала определить зависимость минимального значения *AMDF* от отношения сигнал – шум $y_{\min}$(с/ш). Построить график зависимости.
2. Анализируя форму графика *AMDF*, определить минимально возможное отношение сигнал – шум, при котором измерение периода (частоты) основного тона происходит без сбоев.
3. Для найденного отношения сигнал – шум выставить минимально возможное значение порога $r_к$, которое соответствует признаку вокализованности. Зафиксировать «осциллограммы» входного сигнала ИОТ, выходного сигнала ФНЧ и график *AMDF* для данного случая.
4. Зафиксировать график *AMDF* и «осциллограммы» для заданного вокализованного сегмента реального звукового сигнала. Установив найденный порог классификации, определить частоту основного тона и частоту самой мощной форманты (по графику). Измерить значения функции AMDF $y_{\min\ от}$ и $y_{\min\ ф}$ для основного тона и форманты соответственно.
5. Для заданного звукового файла оценить на слух искажения сигнала, «пропущенного» через речевой кодек при значениях классификационного порога, равных найденному, а также больших и меньших его. Найти значение порога для случая минимальных искажений.

Содержание отчета

1. Структурная схема процесса обработки сигнала при исследовании измерителя основного тона.
2. Таблицы и графики определенных зависимостей, «осциллограммы» сигналов.

3. Результаты прослушивания звуковых файлов.
4. Выводы по работе.

Контрольные вопросы

1. Как изменится *AMDF* при увеличении отношения сигнал – шум (сигналом является импульсная последовательность)?
2. К чему приведет чрезмерное уменьшение значения порога классификации?
3. Как изменится график *AMDF* при сужении полосы пропускания фильтра низких частот, установленного перед измерителем основного тона (сигналом является импульсная последовательность)?
4. Почему на вход измерителя основного тона в кодерах сигнал подают через анализирующий фильтр?
5. Как изменится форма графика *AMDF* при увеличении порядка анализирующего фильтра, стоящего перед ним, от 2 до 3?
6. В чем суть корреляционного метода измерения периода основного тона?
7. С какой целью при вычислении *AMDF* используется нормирующий коэффициент?
8. С какой целью при определении периода основного тона вводится измерительный порог?
9. К чему приведет исключение фильтра нижних частот из состава измерителя основного тона?
10. Работу какого органа речи характеризует частота основного тона?

3. Лабораторная работа

Исследование кодера параметров речевого сигнала

Цель работы

Исследовать зависимость искажений синтезированного речевого сигнала от количества разрядов двоичного квантования параметров сигнала.

Краткие теоретические сведения

Назначением кодера параметров является такая «упаковка» параметров, которая обеспечивает наивысшую степень сжатия сигнала при его допустимых искажениях. «Упаковка» достигается квантованием параметра (представлением его малоразрядным двоичным числом) с предварительным нелинейным преобразованием.

В результате анализа сигнала - кодирования - сегмент сигнала может быть достаточно точно представлен следующим набором параметров: энергией E (громкостью звука), огибающей спектра, представленной набором $\boldsymbol{K}$ – коэффициентов частной корреляции, признаком типа возбуждения синтезирующего фильтра («тон/шум»), а также частотой $F_{от}$ (периодом $T_{от}$) основного тона сигнала возбуждения, которая характеризует тональность звучания. Так как энергия и период могут меняться в очень широких пределах, то удобнее кодировать не само значение параметра, а его логарифм. В этом случае постоянная абсолютная погрешность квантования логарифма параметра приводит к постоянству относительной погрешности квантования самого параметра.

При кодировании коэффициентов частной корреляции $\boldsymbol{K} = \{ K_i \}$ целесообразно предварительно их преобразовать согласно следующему выражению:

$$g_i = \ln \frac{(1 - K_i)}{(1 + K_i)}, \quad \mathrm{i} = (1 \ldots P_a), \tag{3.1}$$

где P_a - количество коэффициентов частной корреляции.

В этом случае значения $|K_i|$, близкие к единице и оказывавшие наибольшее влияние на спектр сигнала при синтезе, при квантовании искажаются в наименьшей степени.

Исследования показали, что с увеличением i уменьшается влияние коэффициента частной корреляции на искажения сигнала, а следовательно, для него допустима большая ошибка квантования. Эксплуатация речевых кодеков показала, что представление любого параметра двоичным числом, содержащим восемь и более разрядов, не приводит к существенным искажениям сигнала при синтезе. Наиболее вероятно нахождение каждого параметра в следующих пределах:

$$\ln E \approx [0\ldots15];\ \ln F \approx [3{,}9\ldots6{,}21];\ g \approx [+31\ldots-31]. \qquad (3.2)$$

Исследования в лабораторной работе проводятся с использованием приведенных числовых данных .

Схема обработки сигналов в лабораторной работе

Структурная схема обработки сигналов в лабораторной работе приведена на рис. 3.1.

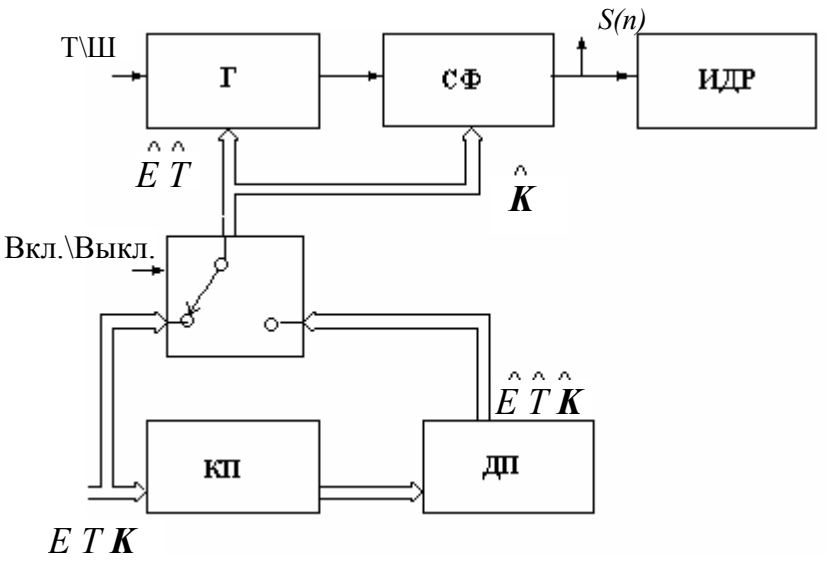


Рис. 3.1

Задаваемые параметры: энергия E, период основного тона $T_{от}$ и набор коэффициентов частной корреляции $\boldsymbol{K}$ для огибающей спектра сигнала - сначала поступают в кодер параметров КП, где они кодируются с заданной точностью, задаваемой числом разрядов двоичного числа (битов) кода параметра. На генератор Г сигнала возбуждения синтезирующего фильтра СФ и на сам фильтр в зависимости от состояния ключевого элемента поступает либо исходный набор параметров, либо этот же набор после кодека параметров.

Степень искажения сигнала из-за погрешности кодирования оценивается путем визуального сравнения «осциллограмм» сигнала, синтезированного либо по исходному набору параметров, либо по набору после кодека. Объективная оценка искажений определяется измерителем нормированной оценки дисперсии разности (ИДР) по формуле

$$D_{\text{н}}(n) = \frac{\sum_{i=1}^{N-1}\left(\hat{S}_i(n) - S_i\right)^2}{\sum_{i=1}^{N-1} S_i^2}, \qquad (3.3)$$

где S_i и $\hat{S}_i(n)$ - выборки сигналов, синтезированные по исходному и прошедшему через кодек (с n–разрядным квантователем) набору параметров, соответственно; N - число выборок в сегменте сигнала. Степень искажений параметра, возникающих при его квантовании, можно оценить также относительной погрешностью квантования:

$$\delta(n) = \frac{\hat{P}(n) - P}{P} 100\%,$$

где P и $\hat{P}(n)$ – значения параметра (исходного и прошедшего через квантователь с числом разрядов n).

Порядок выполнения работы

1. Изменяя количество n разрядов квантователя от 2 до 8 поочередно для каждого параметра, определить для заданного перечня параметров звуков с

помощью «осциллограмм» синтезированного сигнала минимальное количество разрядов $n_{\min}$, при котором искажения сигнала мало заметны.

Для каждого значения n фиксировать величины погрешностей $\delta(n)$ декодирования параметра (в процентах) и нормированной дисперсии $D_{н}(n)$ разности между сигналами, синтезированными с максимальной точностью и точностью, определяемой текущими значениями n.

Зафиксировать «осциллограммы» для $n_{\min}$ и исходного набора параметров.

2. Построить графики зависимостей $D_{н}(n)$, $\delta(n)$ (можно использовать графики, построенные автоматически в ходе лабораторной работы).

3. По зафиксированным «осциллограммам» и построенным графикам определить минимальное суммарное количество $n_{\sum\min}$ разрядов квантования для исследуемых параметров (по каждому звуку). Выбрать максимальное из всех анализируемых случаев звуков значение

$$\max\{n_{\sum\min}\} = n^{\max}_{\sum\min}.$$

Зафиксировать «осциллограммы» заданных звуков для найденного значения $n^{\max}_{\sum\min}$.Оценить расчетным путем изменения коэффициента сжатия речевого сигнала при изменении количества разрядов n.

4. Прослушать контрольную запись после прохождения ее через кодек с параметрами квантования, определенными в предыдущем пункте, и оценить искажения сигнала на слух по критериям разборчивости и узнаваемости диктора по голосу. Прослушать контрольную запись после прохождения ее через кодек при количестве разрядов квантования $n = 8$ для каждого параметра речевого сигнала. Сравнить результаты прослушивания в обоих случаях.

Содержание отчета

1. Схема исследования кодека параметров.
2. Таблицы и соответствующие графики.
3. Изображения «осциллограмм».
4. Выводы по работе.

Контрольные вопросы

1. Почему перед квантованием коэффициентов частной корреляции производится их нелинейное преобразование?
2. Почему значение частоты основного тона перед квантованием логарифмируется?
3. Как наиболее целесообразно следует распределить количество двоичных знаков кода между коэффициентами частной корреляции?
4. Как изменятся параметры кодека, если значение частоты основного тона квантовать без процедуры предварительного логарифмирования?
5. Какие сегменты речевого сигнала требуют большей точности при квантовании их параметров?
6. Для каких групп параметров речевого сигнала целесообразно использование векторного квантования, а для каких – скалярного?
7. Чему равно минимальное количество разрядов квантователя для самого «влиятельного» параметра речевого сигнала?
8. В чем заключается отличие векторного квантования от скалярного?
9. Имеется цифровой канал связи с пропускной способностью 10 кбит/с. Можно ли пропустить через него поток данных речи после LPC-кодера с возбуждением импульсами основного тона, который содержит кодер параметров, представляющий все коэффициенты частной корреляции 50 двоичными разрядами, а частоту основного тона и коэффициент усиления 16 разрядами? Речевой кодер анализирует каждую секунду 40 сегментов сигнала.

4. Лабораторная работа

Исследование *CELP*–кодера речевого сигнала

Цель работы

Исследовать зависимость уровня искажений выходного сигнала *CELP*-кодека от размера кодовой книги и типа звука. Сравнить *CELP*- и *LPC*-кодеки по уровню искажений, вычислительным затратам и по степени сжатия потока данных речи.

Краткие теоретические сведения о кодировании речевого сигнала при возбуждении синтезирующего фильтра кодом

Крупным недостатком *LPC* (*Linear Prediction Coding*)-кодека с возбуждением синтезирующего фильтра импульсами основного тона является низкое качество синтезируемой речи, которое обусловлено грубостью моделирования сигнала возбуждения (либо шум, либо последовательность импульсов). Разделение всех звуков на вокализованные и невокализованные является грубым, поскольку существуют звуки речи (звуки «Ж», «З»), синтез которых требует наличия в сигнале возбуждения коррелированных шумового и периодического компонентов. Кроме того, в вокализованной речи не всегда удается точно определить период основного тона.

При кодировании речевого сигнала с использованием возбуждения синтезирующего фильтра кодом (*Code Exciting Linear Prediction – CELP*) сигнал возбуждения моделируется более точно – качество синтеза речевого сигнала намного выше. Для этого после анализирующего фильтра, присутствующего в схеме *LPC*-кодека, с частотной характеристикой

$$A(z) = 1 - \sum_{i=1}^{M} a_i z^{-i},$$

где a_i – коэффициенты «кратковременного» предсказания, M – порядок предсказания, включают фильтр «долговременного» анализа (на основе

предсказывающего фильтра - предиктора основного тона). Частотная характеристика фильтра имеет вид:

$$P(z) = 1 - \sum_{i=-1}^{2} \beta_i Z^{-(m+i)},$$

где β_1, β_2, β_3 - коэффициенты предсказания долговременного предиктора, $mT_д$ - период основного тона, $Z^{(m+i)}$ - задержка речевого сигнала на $m+i$ отсчетов, $T_д$ – период дискретизации.

Назначение данного фильтра состоит в устранении корреляции в сигнале остатка предсказания, которую вносят импульсы основного тона. Структура фильтра долговременного анализа такая же, как и у анализирующего фильтра *LPC*–кодера (последний является фильтром кратковременного анализа сигнала). Однако в отличие от анализирующего фильтра, он содержит меньшее число звеньев (в основном — три звена).

Принцип работы фильтра долговременного анализа аналогичен работе анализирующего фильтра и отличается лишь величиной задержки входного сигнала в первом звене фильтра. У анализирующего фильтра она равна периоду дискретизации входного сигнала $T_д$, а у фильтра долговременного анализа - ($T_{от}$ - $T_д$), где $T_{от}$ - период следования импульсов основного тона.

В процессе прохождения сегмента сигнала (кадра) через фильтр долговременного анализа он разбивается на несколько субкадров, далее вычисляются значения трех коэффициентов долговременного предсказания: β_1, β_2 и β_3 для каждого субкадра. Чем больше число субкадров, тем точнее моделируется сигнал возбуждения. Сигнал на выходе фильтра практически лишен корреляционных связей и похож на белый шум.

Для формирования сигнала возбуждения синтезирующего фильтра в структуре кодека имеется фильтр, частотная характеристика которого обратна частотной характеристике фильтра долговременного анализа. Фильтр получил название фильтра долговременного синтеза. Его частотная характеристика имеет вид:

$$N(z) = \frac{1}{P(z)} = \frac{1}{1 - \sum_{i=-1}^{2} \beta_i Z^{-(m+i)}}$$

Для точного восстановления речевого сигнала необходимо на вход синтезирующего фильтра $N(z)$ подать сигнал с выхода фильтра долговременного анализа, который, как указывалось выше, очень близок к белому шуму. Однако описание такого сигнала приводит к появлению большого потока данных. С целью его сокращения субкадры на выходе фильтра долговременного синтеза подвергают векторному квантованию – их заменяют наборами отсчетов белого шума, которые хранятся в так называемой кодовой книге. Количество наборов ограничено, и они пронумерованы. Номер набора (вектора) представляется относительно малоразрядным двоичным числом, что и позволяет сократить объем потока данных. В соответствии с заданным номером векторы поступают на вход фильтра долговременного синтеза.

Кодовые книги бывают двух видов: детерминированные и стохастические. Детерминированная кодовая книга образуется из большого количества образцов речевого сигнала. Векторы параметров в процессе создания книги извлекаются из случайной разговорной речи достаточно большой длительности (30 … 40 мин) на мужских и женских голосах. Кодовые книги, образованные на основе реального речевого сигнала, называются детерминированными.

В отличие от детерминированных кодовых книг существуют стохастические, которые составляются из случайной последовательности с равномерным энергетическим спектром. Стохастические кодовые книги обусловливают меньшую точность синтеза сигнала. Однако преимущество стохастической кодовой книги заключается в том, что ее образование обходится без длительного процесса «обучения» на реальных речевых сигналах.

Для снижения искажений в кодере используется так называемый метод анализа (кодирования) сигнала через его синтез (декодирование). При этом на этапе кодирования производится декодирование, и синтезированный сигнал сравнивается с входным сигналом кодера.

Параметры кодирования настраиваются так, чтобы обеспечить минимум отличий между сравниваемыми сигналами.

Выбор оптимального вектора из кодовой книги

В кодирующем устройстве на вход синтезирующего фильтра подаются векторы из кодовой книги. Сначала подается вектор под номером 1. Путем оптимальной настройки коэффициента *G* усиления оценка дисперсии *E* разности между исходным речевым сигналом *S(n)* и синтезированным речевым сигналом *S′(n)* минимизируется. Оптимальный коэффициент усиления

$$G_{oi} = \frac{R_{ilu}}{R_{iuu}}, \text{ где } R_{ilu} = \sum_{n=0}^{L-1} l_i(n) u_i(n), \quad R_{iuu} = \sum_{n=0}^{L-1} u_i^2(n),$$

где i – номер субкадра, $l_i(n)$ и $u_i(n)$ – остаток предсказания на выходе фильтра долговременного анализа и последовательность шума кодовой книги соответственно, L – число отсчетов в субкадре.

Затем номер (индекс, код) последовательности шума в кодовой книге увеличивается на единицу и снова вычисляется оценка дисперсии. Эта процедура повторяется до тех пор, пока не будет проанализировано все содержание кодовой книги и не будет найдена оптимальная последовательность шума, обеспечивающая минимум *E*. Эта оптимальная последовательность шума, то есть оптимальное возбуждение в виде индекса вектора *N* одновременно с коэффициентом усиления и параметрами кратковременного и долговременного анализирующих фильтров $A(z)$ и $P(z)$ передается на декодер.

В декодере по кодовому слову (индексу вектора) из кодовой книги, которая является точной копией кодовой книги кодера, извлекается соответствующий вектор возбуждения $u_i(n)$. Этот вектор умножается на коэффициент усиления G_{oi}, после чего сигнал $G_{oi}u_i(n)$ поступает на вход долговременного синтезирующего фильтра, а затем на вход фильтра кратковременного синтеза. Передаточная функция последовательности этих фильтров имеет вид:

$$H(z) = \frac{1}{1 - \sum_{i=-1}^{2} \beta_i Z^{-(m+i)}} \quad \frac{1}{1 - \sum_{i=1}^{M} a_i Z^{-i}} = \frac{1}{P(z)A(z)}.$$

Структурная схема исследуемого *CELP*-кодера приведена на рис. 4.1.

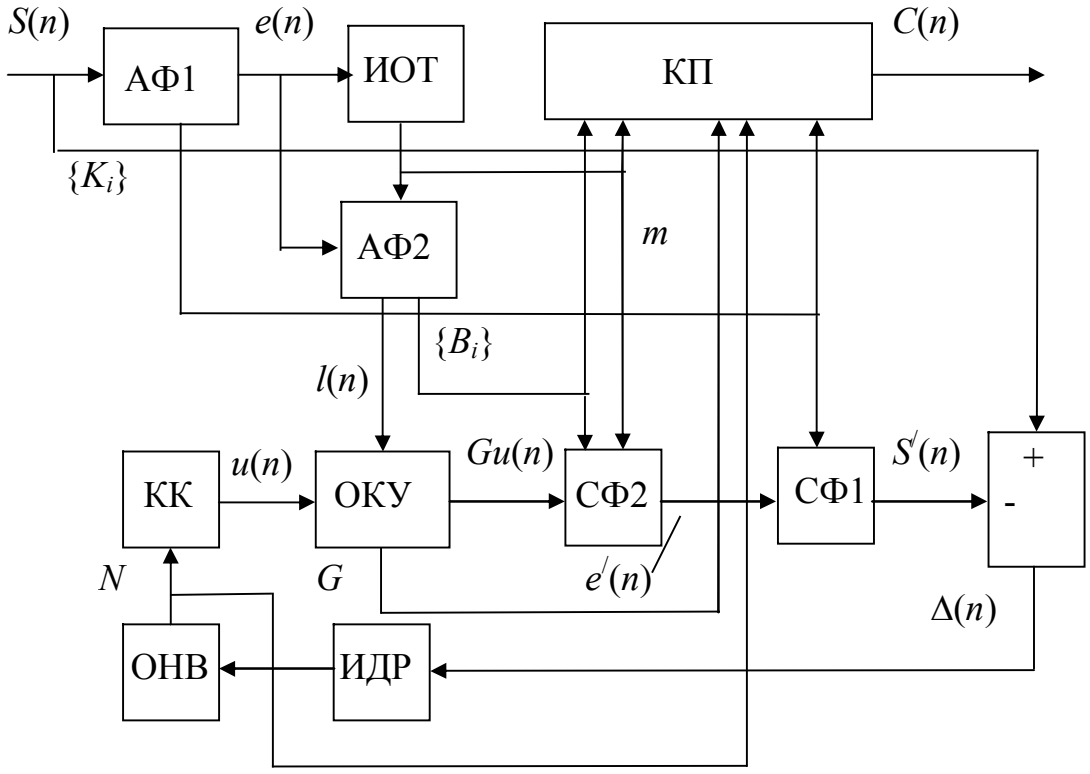


Рис. 4.1

Поток $S(n)$ данных «оцифрованного» с частотой дискретизации 8 кГц речевого сигнала (каждый отсчет сигнала представлен восьмиразрядным двоичным числом) поступает на фильтр кратковременного анализа АФ1 лестничного типа. В процессе анализа кадра сигнала (180 отсчетов) определяется 10 коэффициентов частной корреляции $\{K_i\}$, которые однозначно связаны с коэффициентами предсказания $\{a_i\}$. Сигнал ошибки предсказания $e(n)$ поступает на измеритель основного тона ИОТ, определяющий период основного тона $mT_д$ ($T_д$ – период дискретизации), и фильтр долговременного анализа АФ2. Выходной сигнал $l(n)$ фильтра разбивается на шесть субкадров длительностью 3,75 мс (30 отсчетов речевого сигнала) и используется определителем оптимального коэффициента усиления ОКУ. Коэффициенты частной корреляции $\{B_j\}$ субкадра однозначно связаны с коэффициентами «долговременного» предсказания $\{\beta_j\}$.

Из стохастической кодовой книги КК , содержащей 1024 вектора размером 30, извлекается первый вектор. ОКУ определяет оптимальный коэффициент усиления и формирует аппроксимацию $Gu(n)$ остатка предсказания $l(n)$, которая поступает на «долговременный» синтезирующий фильтр СФ2.

Здесь синтезируется остаток предсказания $e'(n)$, который поступает на «кратковременный» синтезирующий фильтр СФ1. Синтезированный речевой сигнал $S'(n)$ сравнивается с входным речевым сигналом. Найденная разность $\Delta(n)$ поступает на измеритель дисперсии разности ИДР, где определяется дисперсия разности. Определитель номера N вектора ОНВ обеспечивает поиск оптимального вектора, который минимизирует дисперсию.

В кодере параметров речевого сигнала КП формируется поток данных $C(n)$, поступающий в канал связи. На приемной стороне поток данных декодируется декодером параметров, и далее синтезируется речевой сигнал .

Исследования показали, что кодек параметров при допустимых искажениях речевого сигнала может передать всю информацию о коэффициентах кратковременного анализирующего фильтра 41 битом, о периоде основного тона – 7 битами, о коэффициентах фильтра долговременного синтеза – 9 битами, об оптимальном коэффициенте усиления вектора возбуждения – 4 битами, о номере вектора из книги размером 1024 – 10 битами.

Таким образом, для передачи одного кадра потребуется 41 + 7 + (9 + 4 + +10) 6 = 186 битов. Если кадры следуют с частотой 100 Гц, то скорость потока данных на выходе кодера составит 186·100 = 18600 бит/с. Следует отметить, что для передачи данных кратковременного анализа можно использовать векторное квантование с кодовой книгой размером 1024, что требует использования лишь 10 битов вместо 41. Это сокращает поток данных до 15500 бит/с.

Для *LPC*-кодека передача данных о частоте основного тона и признака «тон/шум» требует использования 7 битов, а на передачу коэффициента усиления «тратится» 5 битов. Следовательно, передача данных о кадре сигнала требует 41 + 7 + 5 = 53 бита, что обусловливает скорость передачи данных 5300 бит/с.

Схема обработки речевого сигнала при исследовании CELP–кодека

Схема изображена на рис. 4.2.

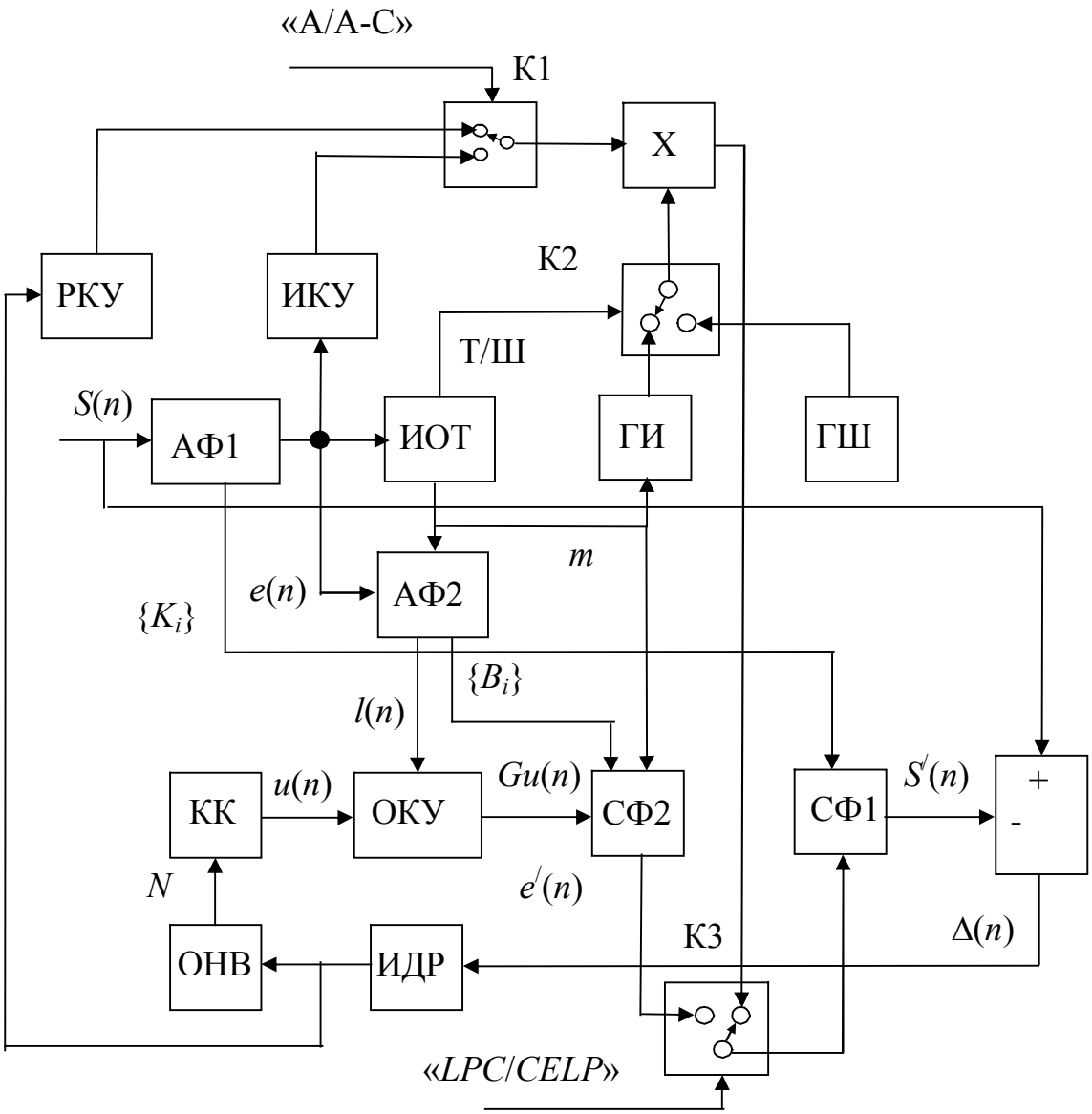


Рис.4.2

Исследование проводится путем сравнения уровня искажений выходного сигнала *CELP*-кодека с искажениями, возникающими в случае с *LPC*-кодеком, использующим возбуждение синтезирующего фильтра импульсами основного тона. Причем рассматривается ситуация, когда определение коэффициента усиления в *LPC*-кодеке производится двумя способами. При первом способе величина коэффициента усиления определяется измерителем коэффициента усиления ИКУ по

энергии остатка $e(n)$ предсказания, а при втором – производится регулировка коэффициента усиления (РКУ) по критерию минимума оценки дисперсии разности $\Delta(n)$ с целью нахождения оптимального его значения. В последнем случае имеет место простейшая реализация метода анализа речевого сигнала через его синтез. Этот метод используется и при *CELP*-кодировании. Использование того или другого способа определяется сигналом «А/А-С» управления ключом К1.

Переключение режима исследования (*LPC* – *CELP*-кодек) осуществляется сигналом «*LPC*/*CELP*» управления ключом К3. Ключ К2 управляется сигналом «тон/шум» («Т/Ш»), он обеспечивает формирование сигнала возбуждения для СФ1 либо с помощью генератора импульсов ГИ, либо с помощью генератора шума ГШ.

Порядок выполнения работы

1. Для заданного звука определить оценку дисперсии разности между входным и выходным сигналами простейшего *LPC*-кодека. Зафиксировать параметры сигналов, а также их «осциллограммы», а также «осциллограммы» сигналов возбуждения и остатка предсказания.

2. Повторить п. 3.1 при определении коэффициента усиления *LPC*-кодека методом «анализ через синтез». Зафиксировать значения оптимального коэффициента усиления, а также модулей максимальных значений сигнала возбуждения и выходного сигнала.

3. Определить зависимость дисперсии разности между входным и выходным сигналами *CELP*-кодека от размера кодовой книги. Зафиксировать данные исследований для максимального и минимального размеров кодовой книги: «осциллограммы» входного и выходного сигналов, а также сигналов остатков предсказания после фильтров «кратковременного» и «долговременного» анализа на выходе кодовой книги, параметры речевого сигнала. Дать оценку роста вычислительных затрат, измеряя время кодирования, при увеличении размера кодовой книги.

Минимальный размер кодовой книги определяется из условия примерного равенства уровней искажений для *CELP*- и *LPC*-кодеков.

4. Повторить пп. 1…3 для второго заданного звука.

5. Сравнить все рассмотренные варианты кодирования речевого сигнала по степени искажений, оцененных по «осциллограммам» и дисперсиям разностей.

Содержание отчета

1. Структурная схема процесса обработки сигнала при исследовании кодека.
2. Данные о результатах исследований: параметры сигналов, графики, «осциллограммы».
3. Выводы по работе: сравнительная характеристика кодеков по искажениям для различных звуков, вычислительным затратам и степени сжатия потока данных.

Контрольные вопросы

1. Чем объясняется меньший уровень искажений *CELP*-кодека по сравнению с *LPC*-кодеком?
2. Почему при оценке уровня искажений выходных сигналов кодеков используют объективные и субъективные критерии?
3. Когда целесообразно использовать стохастические кодовые книги?
4. Из каких соображений устанавливается количество субкадров в кадре?
5. В чем заключаются достоинства и недостатки кодирования речевых сигналов методом «анализ через синтез»?
6. В каких случаях целесообразно использовать *LPC*-кодек?
7. Как зависит степень искажений выходного сигнала *CELP*-кодека от уровня мощности источника шума, используемого при создании стохастической кодовой книги?

5. Лабораторная работа

Исследование системы распознавания голосовых команд на основе теории распознавания образов

Цель работы

Исследование точности распознавания голосовых команд в зависимости от набора их признаков.

Краткие теоретические сведения

Простейшая система распознавания голосовых команд (ГК) включает в себя следующие этапы обработки речевых сигналов: запись сигналов в память; удаление пауз из записи; определение параметров сигнала, «очищенного» от пауз; определение набора признаков формы «траектории» каждого параметра во времени; классификация набора признаков – распознавание речевой команды. Схематически данный процесс представлен на рис. 5.1.

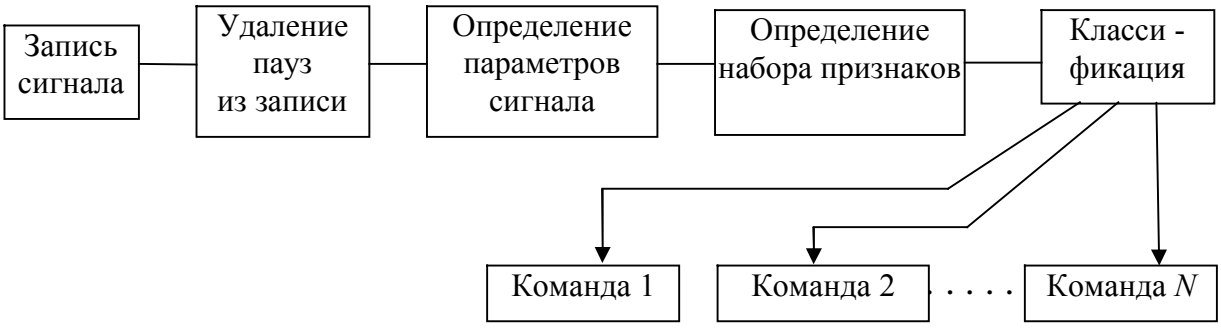
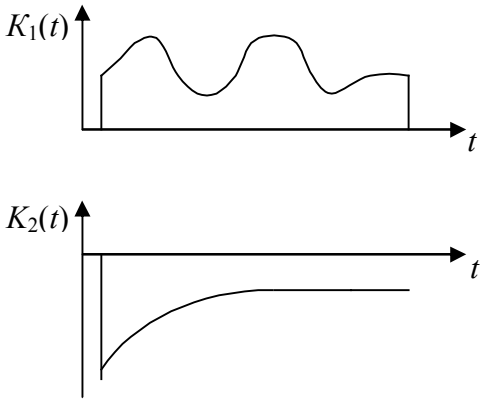


Рис. 5.1

Определение признаков формы траектории требуется для уменьшения зависимости образа команды от скорости и манеры ее произнесения. На рис. 5.2 приведены примеры определения признаков команды по траектории ее параметра во времени.

Каждый набор признаков можно рассматривать как многомерный вектор в некотором пространстве (или в виде точки многомерного пространства). В виду сильной изменчивости произнесения команды ее наборы признаков изменяются при повторе команды, а также при смене диктора. При этом точки многомерного пространства группируются в некоторые области, соответствующие определенным голосовым командам.



Признаки:
1) $K_1(t) > 0$, 2) 2 минимума для $K_1(t)$;

3) $K_2(t) < 0$, 4) $K_2(t)$ – убывающая по модулю функция.

Рис. 5.2

Из-за сильной изменчивости ГК области могут перекрываться, что обусловливает появление ошибок распознавания. Чем больше число команд, тем сложнее разделить области, поэтому приходится повышать размерность пространства – увеличивать количество признаков. Если число ГК очень велико, то каждую ГК сегментируют на отдельные звуки, слоги. Количество звуков в языке ограничено, поэтому, классифицируя сегменты, а не ГК, можно уменьшить число классов и снизить вероятность ошибки распознавания. Из-за неточной сегментации и наличия ошибок в распознавании звуков в состав системы включают лингвистический декодер.

При обучении системы заданному набору ГК определяется положение гиперплоскостей, разделяющих области многомерного пространства. Случай двумерного пространства проиллюстрирован рис. 5.3.

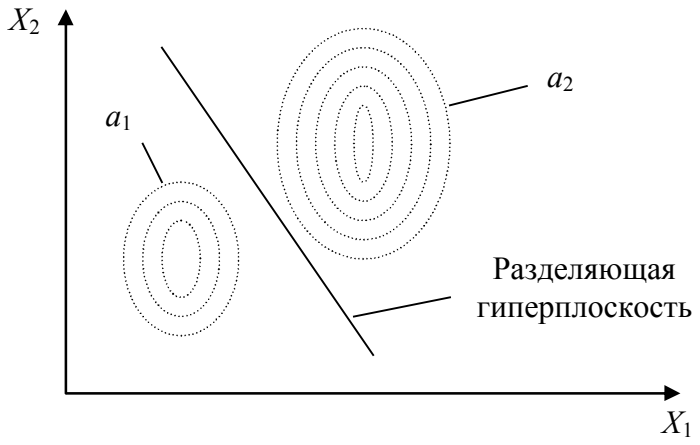


Рис. 5.3

На рис. 5.3 X_1, X_2 – значения признаков, a_1 и a_2 – многомерные области.

Уравнение гиперплоскости:

$$w_1x_1 + w_2x_2 + w_3 = 0.$$

Более короткая запись уравнения в векторном виде:

$$\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X} = 0,$$

где $\boldsymbol{W}^{\boldsymbol{T}} = \{w_1, w_2, w_3\}$ - вектор весовых коэффициентов, $\boldsymbol{X}^{\boldsymbol{T}} = \{x_1, x_2, 1\}$ - вектор признаков.

Если $d(\boldsymbol{X}) = \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X} > 0$, то $\boldsymbol{X} \in a_2$;

$d(\boldsymbol{X}) = \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X} < 0$, то $\boldsymbol{X} \in a_1$.

Здесь $d(\boldsymbol{X})$ является решающей функцией.

Весовой вектор $\boldsymbol{W}$ находится на стадии обучения итерационной процедурой:

$$\boldsymbol{W}(k+1) = \boldsymbol{W}(k) + k^{-1}\boldsymbol{X}(k)[\Gamma(\boldsymbol{X}(k)) - \boldsymbol{W}^T(k)\boldsymbol{X}(k)].$$

Здесь $\boldsymbol{W}(k)$ - весовой вектор на k-м шаге итерационного процесса, $\boldsymbol{X}(k)$ - вектор признаков, предъявляемый на k-й итерации для обучения системы.

$$\Gamma(\boldsymbol{X}) = \begin{cases} 1, \text{ если } X \in a_1; \\ \\ -1, \text{ если } X \notin a_1. \end{cases}$$ - переменная классификации.

Процесс обучения длится до тех пор, пока достаточно большой процент предъявляемых для обучения векторов не будет распознан правильно с использованием определенного весового вектора $\boldsymbol{W}$. Длительность обучения определяется количеством итераций.

Количество признаков голосовой команды определяет размерность пространства. Чем больше признаков, тем проще разделить многомерные области, но при этом увеличивается объем вычислений.

Порядок выполнения работы

1. По предварительно созданной выборке звуковых файлов команд «ноль» и «один», провести «обучение» системы распознавания ГК - определить весовой вектор $\boldsymbol{W}$ для заданного перечня признаков. Зафиксировать число итераций, необходимых для создания вектора, - определить длительность обучения.

2. Провести распознавание нескольких команд «ноль» и «один», фиксируя $|d_{\min}|$ для $d > 0$ и $d < 0$. Оценить вероятность ошибки.

3. Уменьшить число признаков команд сначала на 1, а затем на 2 в соответствии с заданием. Повторить пункты 1, 2.

4. Для полного перечня признаков провести обучение системы по заданной или созданной самостоятельно обучающей выборке (наборы файлов «ноль» и «один»). Провести распознавание набора команд «ноль» и «один». Оценить вероятность ошибки и $|d_{\min}|$ для $d > 0$ и $d < 0$.

Содержание отчета

1. Структурная схема системы распознавания ГК.
2. Таблицы данных по исследованию точности распознавания.
3. Значения весовых векторов $\boldsymbol{W}$, $|d_{\min}|$.
4. Выводы по работе.

Контрольные вопросы

1. Почему системы распознавания речевых сигналов не нашли широкого применения?

2. В каких случаях в системах распознавания речи целесообразно использовать разложение звукового образа слова на отдельные элементы?
3. Почему в общем случае, увеличение количества параметров речевого сигнала увеличивает надежность распознавания?
4. Что общего между системами эффективного кодирования и распознавания речевых сигналов?
5. Как с позиций теории распознавания образов объяснить факт увеличения ошибки распознавания при росте числа распознаваемых голосовых команд?
6. Какие основные факторы определяют точность распознавания голосовых команд?
7. В каких случаях целесообразно осуществлять распознавание голосовой команды на основе разложения ее на отдельные звуки или сочетания звуков?
8. Какими факторами определяется длительность обучения системы?
9. Каким образом выбор системы параметров речевого сигнала может повлиять на точность распознавания?
10. Какими недостатками обладает рассмотренный в лабораторной работе метод формирования набора признаков голосовой команды?

6. Лабораторная работа

Исследование системы распознавания голосовых команд на основе теории скрытых марковских процессов

Цель работы

Определить параметры моделей скрытых марковских процессов – эталонов заданных голосовых команд, надежность распознавания команд с использованием полученных эталонов, проанализировать причины ошибок распознавания.

Краткие теоретические сведения

При автоматическом распознавании голосовых команд (ГК) происходит сопоставление временной последовательности наборов (векторов) параметров речевого сигнала (РС) с эталонами распознаваемых команд. Эталон, с наибольшей вероятностью соответствующий произнесению ГК, является результатом распознавания.

Распространенным подходом к построению систем автоматического распознавания речи (АРР) является использование эталонов в виде моделей скрытых марковских процессов (МСМП). В литературе часто упоминается метод построения МСМП, основанный на итерационной процедуре Баума – Уэлча с применением алгоритма Витерби.

Рассмотрим систему (рис. 6.1), про которую можно сказать, что она находится в одном из набора $\boldsymbol{S} = \{S_1, S_2,\ldots S_N\}$ состояний («Н», «К» - начальное и конечное состояния).

В каждый из дискретных моментов времени $t = 1, 2,\ldots$ система совершает переход из одного состояния в другое, находящееся правее данного, согласно набору вероятностей, связанных с данным состоянием.

Вероятности a_{ij} перехода из i–го в j–е состояние в данный момент времени составляют матрицу переходных вероятностей. Обозначим ее как $\boldsymbol{A}$. Существует также набор вероятностей возникновения i–го начального состояния $\boldsymbol{\pi} = \{\pi_i\}$.

Допустим, что состояния некоторой системы не наблюдаемы, то есть нет возможности точно определить состояние q_t, в котором данная система

находится в момент времени t. Допустим, что существует некоторый набор из M символов наблюдения $\boldsymbol{V} = \{v_1, v_2, \ldots, v_M\}$, такой, что в каждом из своих состояний система порождает какой-либо символ из $\boldsymbol{V}$.

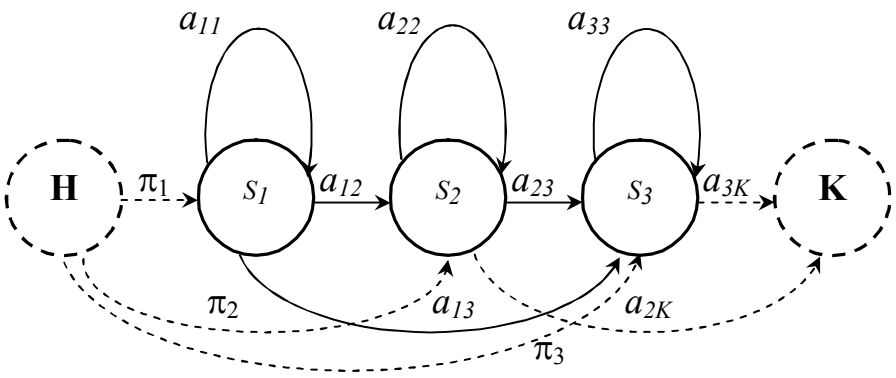


Рис.6.1

Появление символа наблюдения описывается вероятностями:

$$b_j(k) = P[o_t = v_k \mid q_t = S_j], \qquad 1 \le k \le M,$$

где o_t – символ из множества $\boldsymbol{V} = \{v_1, v_2, \ldots, v_M\}$, наблюдаемый в момент времени t, j – номер состояния.

Аналогично матрице переходных вероятностей определяется матрица распределения вероятностей появления символов наблюдения $\boldsymbol{B} = \{b_j(k)\}$. Появление символов наблюдения зависит только от состояния, порождающего символ. Если вместо символов наблюдаются аналоговые значения x, то $b_j(x)$ – плотность распределения вероятностей появления x.

В общем случае наблюдается не одно значение x , а некоторый набор $\boldsymbol{X}$ (вектор) значений (например набор параметров речевого сигнала). В этом случае $b_j(\boldsymbol{X})$ – совместная плотность вероятностей набора значений. При наблюдении совокупности параметров речевого сигнала часто для снижения объема вычислений считают, что корреляционная матрица параметров является диагональной, и каждый параметр описывается гауссовым законом распределения вероятностей.

Характеристики рассматриваемой системы полностью определяются набором $\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ — параметров модели скрытого марковского процесса.

Будем считать, что каждой произнесенной команде соответствует система с конечным числом состояний (модель команды). Из каждого состояния возможны переходы трёх типов: возвратный переход, переход к

следующему состоянию и переход скачком через одно состояние. Наличие обратных и скачкообразных переходов в МСМП обеспечивает случайное изменение длины цепочки состояний во времени, МСМП таким образом подстраивается к темпу речи. Всякий раз, когда происходит переход в очередное состояние S_h, образуется выходной вектор $\boldsymbol{V}$ параметров сигнала, который считается случайным гауссовским вектором со средним значением $\boldsymbol{m}_h$ и среднеквадратическим отклонением $\boldsymbol{\delta}_h$, корреляционная матрица параметров – диагональная. Все эти параметры определяются статистическим методом путём анализа большого количества различных вариаций произнесения каждого слова.

Пусть получена некоторая последовательность наблюдаемых векторов (или символов) $\boldsymbol{O} = \{o_1, o_2, \ldots, o_T\}$ во времени. Необходимо определить, с какой вероятностью данная последовательность соответствует некоторой МСМП. Распознавание голосовых команд заключается в нахождении такой МСМП из имеющегося перечня моделей, которая с наибольшей вероятностью соответствует наблюдаемой последовательности векторов. Точное решение этой задачи требует больших вычислительных затрат. На практике используют ее приближенное решение путем нахождения оптимальной последовательности состояний МСМП во времени, которая с наибольшей вероятностью соответствует наблюдаемой последовательности векторов. Нахождение оптимальной последовательности осуществляется с помощью алгоритма Витерби.

Рассмотрим работу алгоритма на конкретном примере. Пусть имеется модель скрытого марковского процесса, которая представлена на рис. 6.1.

Пусть имеется набор из четырех символов наблюдения $\boldsymbol{V} = \{v_1, v_2, v_3, v_4\}$, $b_i(v_j)$ – вероятность появления символа v_j для состояния S_i.

Наблюдается набор символов $\boldsymbol{O} = \{o_1 = v_2, o_2 = v_4, o_3 = v_2, o_4 = v_1, o_5 = v_1\}$ для моментов времени: t_1, t_2, t_3, t_4, t_5. Процесс смены состояний модели в моменты времени: t_1, t_2, t_3, t_4, t_5 отражается диаграммой, представленной на рис. 6.2.

Определим вероятности появления символа v_2 в момент времени t_1 на этапе инициализации МСМП для каждого из возможных состояний в данный момент времени.

$$\delta_1(1) = \pi_1 b_1(o_1) = \pi_1 b_1(v_2), \quad \delta_1(2) = \pi_2 b_2(o_1) = \pi_2 b_2(v_2), \quad \delta_1(3) = \pi_3 b_3(o_1) = \pi_3 b_3(v_2).$$

Для момента времени t_2 вероятности d_2 появления последовательности двух состояний: q_1, $q_{2,}$, которые описываются следующими выражениями.

$$d'_2(S_1) = \delta_1(1)a_{11}b_1(o_2),$$

$\underline{d'_2(S_2) = \delta_1(1)a_{12}b_2(o_2)}$ – максимальная вероятность, если $q_2 = S_2$,

$$d''_2(S_2) = \delta_1(2)a_{22}b_2(o_2),$$

$\underline{d'_2(S_3) = \delta_1(1)a_{13}b_3(o_2)}$ – максимальная вероятность, если $q_2 = S_3$,

$$d''_2(S_3) = \delta_1(2)a_{23}b_3(o_2),$$

$$d'''_2(S_3) = \delta_1(3)a_{33}b_3(o_2).$$

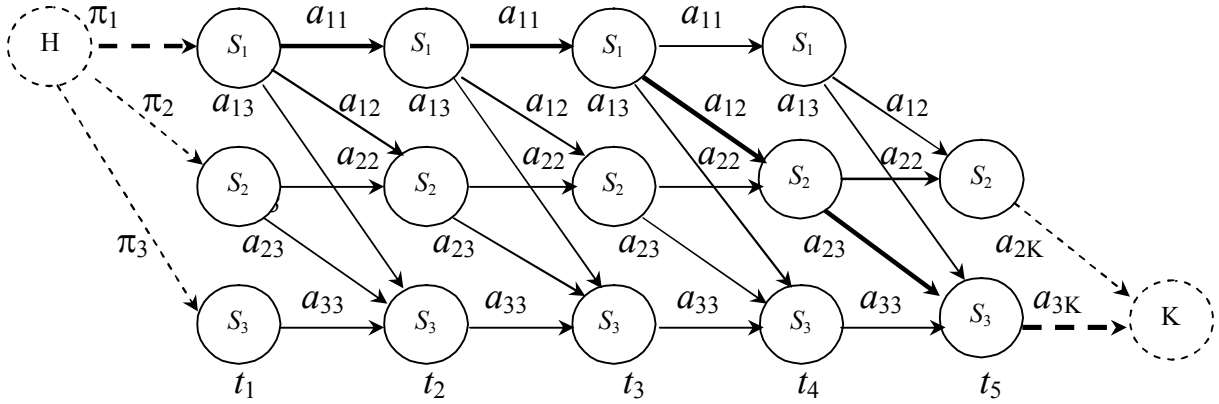


Рис. 6.2

Пусть $d'_2(S_2) > d''_2(S_2)$, а $d'_2(S_3) > d''_2(S_3)$ и $d'_2(S_3) > d'''_2(S_3)$.
Следовательно, максимальные вероятности: для $q_2 = S_1$, $q_2 = S_2$, $q_2 = S_3$ – равны соответственно: $\delta_2(1) = d'_2(S_1)$, $\delta_2(2) = d'_2(S_2)$, $\delta_2(3) = d'_2(S_3)$.

Элементы массива $\psi_2(j)$ (j – номер состояния в момент времени t_2) значений состояния q_1, обеспечивающих максимум вероятности соответствия символов наблюдения o_1, o_2 последовательности q_1q_2 , принимают следующие значения:

$$\psi_2(1) = S_1,\ \psi_2(2) = S_1,\ \psi_2(3) = S_1.$$

То есть максимальные вероятности последовательности состояний q_1q_2 имеют место, если $q_1 = S_1$.

Определим вероятности для момента времени t_3.

$\underline{d'_3(S_1) = \delta_2(1)a_{11}b_1(o_3)}$, $d'_3(S_2) = \delta_2(1)a_{12}b_2(o_3)$, $\underline{d''_3(S_2) = \delta_2(2)a_{22}b_2(o_3)}$,

$d'_3(S_3) = \delta_2(1)a_{13}b_3(o_3)$, $\underline{d''_3(S_3) = \delta_2(2)a_{23}b_3(o_3)}$, $d'''_3(S_3) = \delta_2(3)a_{33}b_3(o_3)$.

Подчеркнутые выражения соответствуют максимальным вероятностям $d_3(S_j)$ появления последовательности состояний q_1 - q_2 - $q_3 = S_j$.

Тогда

$$\delta_3(1) = d'_3(S_1), \delta_3(2) = d''_3(S_2), \quad \delta_3(3) = d'_3(S_2).$$

Соответствующие элементы массива:

$$\psi_3(1) = S_1, \quad \psi_3(2) = S_2, \quad \psi_3(3) = S_2.$$

Применяя аналогичные рассуждения для t_4, получаем:

$\underline{d'_4(S_1) = \delta_3(1)a_{11}b_1(o_4)}$ $\quad$ $d'_4(S_3) = \delta_3(1)a_{13}b_3(o_4)$

$\underline{d'_4(S_2) = \delta_3(1)a_{12}b_2(o_4)}$ $\quad$ $d''_4(S_3) = \delta_3(2)a_{23}b_3(o_4)$

$d''_4(S_2) = \delta_3(2)a_{22}b_2(o_4)$ $\quad$ $\underline{d'''_4(S_3) = \delta_3(3)a_{33}b_3(o_4)}$.

$$\delta_4(1) = d'_4(S_1), \quad \delta_4(2) = d'_4(S_2), \quad \delta_4(3) = d'''_4(S_3).$$

Соответствующий массив

$$\psi_4(1) = S_1, \quad \psi_4(2) = S_1, \quad \psi_4(3) = S_3.$$

В момент времени t_5 имеем:

$d'_5(S_2) = \delta_4(1)a_{12}b_2(o_5)$, $\quad$ $\underline{d''_5(S_3) = \delta_4(2)a_{23}b_3(o_5)}$,

$\underline{d''_5(S_2) = \delta_4(1)a_{22}b_2(o_5)}$, $\quad$ $d'''_5(S_3)=\delta_4(3)a_{33}b_3(o_5)$.

$d'_5(S_3) = \delta_4(1)a_{13}b_3(o_5)$,

$\delta_5(1) = d''_5(S_2)$, $\quad \delta_5(2) = d''_5(S_3)$ и $\psi_5(1) = S_2$, $\quad \psi_5(2) = S_2$.

В конечный момент времени:

$d''_к(S_2) = \delta'_5(2)a_{2K}$, $\quad$ $d''_к(S_3) = \underline{\delta'_5(3)a_{3K}}$. $\quad \delta_к(3) = d_к(S_3)$ и $\psi_к(3) = S_3$.

Восстанавливаем, начиная с конечного состояния, оптимальную последовательность состояний, обеспечивающую максимальную вероятность:

$$S_1 \leftarrow S_1 \leftarrow S_1 \leftarrow S_2 \leftarrow S_3$$

$$q_1 \quad q_2 \quad q_3 \quad q_4 \quad q_5 .$$

Эта последовательность на рис. 6.2 выделена утолщенными линиями.

Вероятность появления оптимальной последовательности состояний $\mathbf{Q} = \{S_1 - S_1 - S_1 - S_2 - S_3\}$ для набора наблюдений $\boldsymbol{O}$

$$P(\boldsymbol{O}, \boldsymbol{Q}) = \pi_1 b_1(o_1)a_{11}b_1(o_2)a_{11}b_1(o_3)a_{12}b_2(o_4)a_{23}b_3(o_5)a_{3K}.$$

Определение параметров МСМП

Правильность распознавания голосовых команд зависит от того, насколько точно команде соответствует ее эталон – МСМП.

Для формирования МСМП (обучения системы распознавания новой команде) часто используется итерационный алгоритм Баума - Уэлча. До начала итерационного процесса обучения необходимо получить начальную модель команды. Ее можно получить, выбирая из всех вариантов произнесения команд наиболее типичный, и по нему определить начальные значения векторов параметров. Также необходимо определить число стационарных состояний МСМП. Процесс трудно автоматизировать, и он выполняется вручную.

После определения начальной модели весь процесс обучения происходит полностью в автоматическом режиме. Необходимо только отобрать записи команды, которые будут участвовать в процессе обучения. Процесс обучения использует алгоритм Витерби и начальную модель команды. Определяются оптимальные последовательности состояний модели для каждого варианта произнесения команды. По полученным последовательностям корректируются параметры модели. Далее весь процесс повторяется до тех пор, пока изменения параметров модели не станут минимально допустимыми. Рассмотрим данный процесс на конкретном примере.

До начала процесса обучения необходимо задаться начальной моделью голосовой команды. Для определения её параметров требуется проанализировать весь имеющийся звуковой материал с целью выбора наиболее типичной реализации голосовой команды. Для нее определяется массив параметров (оценки математических ожиданий – m, равные значениям параметров РС при типичном произнесении), которые и используются на начальном этапе обучения. За начальные значения среднеквадратических отклонений (δ) можно принять значения, в несколько раз меньшие значений самих параметров.

Задаемся следующими значениями вероятностей переходов между состояниями: 0,5 - для перехода в следующее состояний (p_2); 0,25 – для повторов (p_1) состояний и «прыжков» через состояние (p_3). Вероятность перехода из начального состояния в первое (p'_1) примем 0,5, для перехода во второе $p'_2 = 0{,}3$ и для третьего $p'_3 = 0{,}2$.

Пусть имеется два варианта произнесения команды, которые характеризуются последовательностями векторов параметров $\boldsymbol{V} = \{v_1,\ldots, v_6\}$ и $\boldsymbol{W} = \{w_1,\ldots, w_3\}$, соответственно. Каждый вектор – это набор из шести параметров k_i или l_i (i = 1, 2,…, 6).

I вариант

$v_1\{k_1^{t1}, k_2^{t1}, k_3^{t1}, k_4^{t1}, k_5^{t1}, k_6^{t1}\}$
$v_2\{k_1^{t2}, k_2^{t2}, k_3^{t2}, k_4^{t2}, k_5^{t2}, k_6^{t2}\}$
$v_3\{k_1^{t3}, k_2^{t3}, k_3^{t3}, k_4^{t3}, k_5^{t3}, k_6^{t3}\}$
$v_4\{k_1^{t4}, k_2^{t4}, k_3^{t4}, k_4^{t4}, k_5^{t4}, k_6^{t4}\}$

II вариант

$w_1\{l_1^{t1}, l_2^{t1}, l_3^{t1}, l_4^{t1}, l_5^{t1}, l_6^{t1}\}$
$w_2\{l_1^{t2}, l_2^{t2}, l_3^{t2}, l_4^{t2}, l_5^{t2}, l_6^{t2}\}$
$w_3\{l_1^{t3}, l_2^{t3}, l_3^{t3}, l_4^{t3}, l_5^{t3}, l_6^{t3}\}$.

Применим алгоритм Витерби для нахождения последовательности состояний для первого произнесения. Считаем параметры сигнала независимыми и распределенными по гауссовому закону.

Вычислим плотность вероятности того, что первый вектор параметров сигнала порожден первым состоянием.

$$P_{S1}^{(t1)} = \prod_{i=1}^{6} \frac{1}{\delta_i^{(1)}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{k_i^{(t1)} - m_i^{(1)}}{\delta_i^{(1)}}\right)^2\right\},$$

где $m_i^{(j)}$ – начальное значение математического ожидания для i-го параметра и j-го состояния МСМП, $\delta_i^{(j)} = 0{,}2m_i^{(j)}$ – начальное значение среднеквадратического отклонения.

Для второго состояния:

$$P_{S2}^{(t1)} = \prod_{i=1}^{6} \frac{1}{\delta_i^{(2)}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{k_i^{(t1)} - m_i^{(2)}}{\delta_i^{(2)}}\right)^2\right\}.$$

Аналогично определяется плотность вероятности для третьего состояния.

Определим плотность вероятности того, что второй вектор сигнала порожден состоянием S_1:

$$P_{S1}^{(t2)} = \prod_{i=1}^{6} \frac{1}{\delta_i^{(1)}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{k_i^{(t2)} - m_i^{(1)}}{\delta_i^{(1)}}\right)^2\right\}.$$

Аналогично для состояний: S_2 и S_3 получаем плотности вероятностей: $P_{S2}^{(t2)}$ и $P_{S3}^{(t2)}$.

Вычислим, какова плотность вероятности того, что первые два отсчета сигнала соответствуют последовательностям:

$$S_1 \to S_1,\ S_1 \to S_2,\ S_2 \to S_2,\ S_2 \to S_3,\ S_3 \to S_3.$$

Найдем:

$$\left.\begin{array}{l} P_{S1S1} = P_{S1}^{t1} p_1 P_{S1}^{t2} \\ P_{S1S2} = P_{S1}^{t1} p_2 P_{S2}^{t2} \\ P_{S2S2} = P_{S2}^{t1} p_1 P_{S2}^{t2} \end{array}\right\} P_{S2S2} > P_{S1S2}, \qquad \left.\begin{array}{l} P_{S1S3} = P_{S1}^{t1} p_3 P_{S3}^{t2} \\ P_{S2S3} = P_{S2}^{t1} p_2 P_{S3}^{t2} \\ P_{S3S3} = P_{S3}^{t1} p_1 P_{S3}^{t2} \end{array}\right\} P_{S2S3} > P_{S1S3} > P_{S3S3}.$$

Сохраним в массиве P_{S1S1}; P_{S2S2}; P_{S2S3}.

Определим плотность вероятности для случая, когда третий вектор сигнала порожден состоянием S_1:

$$P_{S1}^{(t3)} = \prod_{i=1}^{6} \frac{1}{\delta_i^{(1)}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{k_i^{(t3)} - m_i^{(1)}}{\delta_i^{(1)}}\right)^2\right\}.$$

Для состояний S_2 и S_3 аналогично получаем: P_{S2}^{t3} и P_{S3}^{t3}.

Вычислим плотность вероятности соответствия первых трех отсчетов сигнала последовательностям:

$$S_1 \to S_1 \to S_1,\ S_1 \to S_1 \to S_2,\ S_2 \to S_2 \to S_2,\ S_2 \to S_2 \to S_3,\ S_1 \to S_1 \to S_3,$$
$$S_2 \to S_3 \to S_3.$$

Найдем:

$$\left.\begin{array}{l} P_{S1S1S1} = P_{S1S1} p_1 P_{S1}^{t3} \\ P_{S1S1S2} = P_{S1S1} p_2 P_{S2}^{t3} \\ P_{S2S2S2} = P_{S2S2} p_1 P_{S2}^{t3} \end{array}\right\} P_{S1S1S2} > P_{S2S2S2}, \qquad \left.\begin{array}{l} P_{S2S2S3} = P_{S2S2} p_2 P_{S3}^{t3} \\ P_{S1S1S3} = P_{S1S1} p_3 P_{S3}^{t3} \\ P_{S2S3S3} = P_{S2S3} p_1 P_{S3}^{t3} \end{array}\right\} P_{S1S1S3} > P_{S2S2S3} > P_{S2S3S3}.$$

Сохраним P_{S1S1S1}, P_{S1S1S2}, P_{S1S1S3}.

Четвертому временному отсчету первое состояние не может соответствовать, так как в конечное состояние переходы возможны лишь из второго и третьего состояний (см. рис. 6.1).

Определим, что четвертый вектор порожден состоянием S_2:

$$P_{S2}^{(t4)} = \prod_{i=1}^{6} \frac{1}{\delta_i^{(2)}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{k_i^{(t4)} - m_i^{(2)}}{\delta_i^{(2)}}\right)^2\right\}.$$

Аналогично определяется плотность вероятности для состояния S_3.

Вычислим вероятность соответствия четырех отсчетов сигнала последовательностям:

$S_1 \to S_1 \to S_1 \to S_2$, $S_1 \to S_1 \to S_2 \to S_2$, $S_1 \to S_1 \to S_1 \to S_3$,
$S_1 \to S_1 \to S_2 \to S_3$, $S_1 \to S_1 \to S_3 \to S_3$.

Найдем:

$$\left.\begin{aligned} P_{S1S1S1S2} &= P_{S1S1S1} p_2 P_{S2}^{t4} \\ P_{S1S1S2S2} &= P_{S1S1S2} p_1 P_{S2}^{t4} \end{aligned}\right\} \quad P_{S1S1S2S2} > P_{S1S1S1S2},$$

$$\left.\begin{aligned} P_{S1S1S1S3} &= P_{S1S1S1} p_3 P_{S3}^{t4} \\ P_{S1S1S2S3} &= P_{S1S1S2} p_2 P_{S3}^{t4} \\ P_{S1S1S3S3} &= P_{S1S1S3} p_1 P_{S3}^{t4} \end{aligned}\right\} \quad P_{S1S1S2S3} > P_{S1S1S1S3} > P_{S1S1S3S3}.$$

Вычислим вероятность соответствия последовательностей состояний заданному сигналу.

$$P_{S1S1S2S2}\,0{,}3 < P_{S1S1S2S3}\,0{,}5.$$

Окончательно имеем последовательность: $S_1 \to S_1 \to S_2 \to S_3$.

Аналогично, используя алгоритм Витерби, найдем последовательность состояний для второго варианта произнесения команды.

Допустим: $S_1 \to S_2 \to S_3$.

Определим вероятности переходов между состояниями в марковской модели (см. рис. 6.1).

$$p_{11} = \frac{N(S_1 \to S_1)}{N(S_1 \to S_1) + N(S_1 \to S_2) + N(S_1 \to S_3)} = \frac{1}{1+2+0} = \frac{1}{3};$$

$$p_{12} = \frac{N(S_1 \to S_2)}{N(S_1 \to S_1) + N(S_1 \to S_2) + N(S_1 \to S_3)} = \frac{2}{1+2+0} = \frac{2}{3};$$

$$p_{13} = \frac{N(S_1 \to S_3)}{N(S_1 \to S_1) + N(S_1 \to S_2) + N(S_1 \to S_3)} = \frac{0}{1+2+0} = 0;$$

$$p_{22} = \frac{N(S_2 \to S_2)}{N(S_2 \to S_2) + N(S_2 \to S_3)} = \frac{0}{0+2} = 0;$$

$$p_{23} = \frac{N(S_2 \to S_3)}{N(S_2 \to S_2) + N(S_2 \to S_3)} = \frac{2}{0+2} = 1;$$

$$p_{33} = \frac{N(S_3 \to S_3)}{N(S_3 \to S_3)} = \frac{0}{0} = 0;$$

$$p'_1 = \frac{N(H \to S_1)}{N(H \to S_1) + N(H \to S_2) + N(H \to S_3)} = \frac{2}{2+0+0} = 1;$$

$$p'_2 = \frac{N(H \to S_2)}{N(H \to S_1) + N(H \to S_2) + N(H \to S_3)} = \frac{0}{2+0+0} = 0;$$

$$p'_3 = \frac{N(H \to S_3)}{N(H \to S_1) + N(H \to S_2) + N(H \to S_3)} = \frac{0}{2+0+0} = 0;$$

$$p_{2K} = \frac{N(S_2 \to K)}{N(S_2 \to S_2) + N(S_2 \to S_3) + N(S_2 \to K)} = \frac{0}{0+2+0} = 0;$$

$$p_{3K} = \frac{N(S_3 \to K)}{N(S_3 \to K) + N(S_3 \to S_3)} = \frac{2}{2+0+0} = 1,$$

где N – количество соответствующих переходов.

Определим значения некоторых оценок математических ожиданий:

$$m_1^{(S1)} = \frac{k_1^{(t1)} + k_1^{(t2)} + l_1^{(t1)}}{3}; \qquad m_2^{(S1)} = \frac{k_2^{(t1)} + k_2^{(t2)} + l_2^{(t1)}}{3};$$

$$m_3^{(S1)} = \frac{k_3^{(t1)} + k_3^{(t2)} + l_3^{(t1)}}{3}; \quad m_1^{(S2)} = \frac{k_1^{(t3)} + l_1^{(t3)}}{2}; \quad m_2^{(S2)} = \frac{k_2^{(t3)} + l_2^{(t3)}}{2};$$

$$m_3^{(S2)} = \frac{k_3^{(t3)} + l_3^{(t3)}}{2}; \qquad m_4^{(S3)} = \frac{k_4^{(t4)} + l_4^{(t4)}}{2};$$

$$m_5^{(S3)} = \frac{k_5^{(t4)} + l_5^{(t4)}}{2}; \quad m_6^{(S3)} = \frac{k_6^{(t4)} + l_6^{(t4)}}{2}.$$

Определим значения некоторых оценок среднеквадратических отклонений:

$$\delta_1^{(S1)} = \sqrt{\frac{(k_1^{(t1)} - m_1^{(S1)})^2 + (k_1^{(t2)} - m_1^{(S1)})^2 + (l_1^{(t1)} - m_1^{(S1)})^2}{3-1}};$$

$$\delta_2^{(S1)} = \sqrt{\frac{(k_2^{(t1)} - m_2^{(S1)})^2 + (k_2^{(t2)} - m_2^{(S1)})^2 + (l_2^{(t1)} - m_2^{(S1)})^2}{3-1}}.$$

После определения новых значений параметров модели весь процесс повторяется, пока параметры не станут инвариантными относительно повторений процесса.

Схема обработки речевых сигналов при выполнении лабораторной работы

Схема представлена на рис. 6.3.

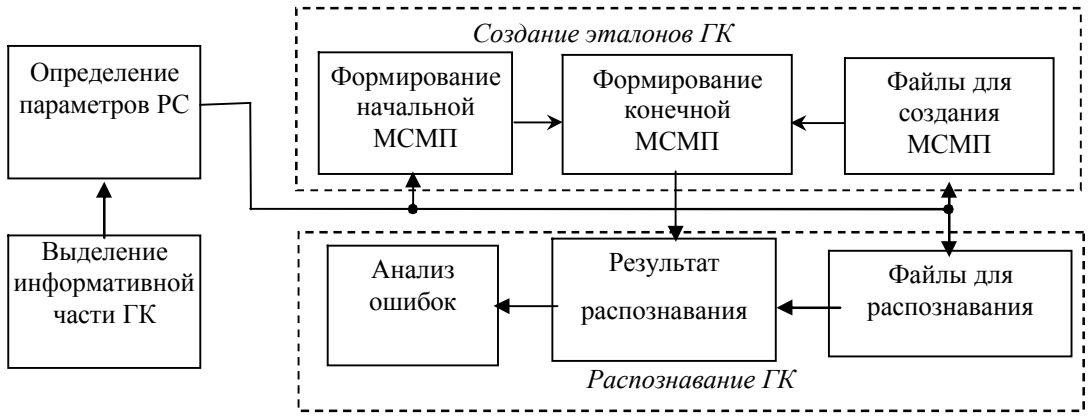


Рис. 6.3

Сначала анализируются звуковые файлы из заданной директории (возможна звукозапись новых файлов с автоматическим присвоением им имен, в которых содержится информация об условиях записи: коды команды, диктора и аппаратуры записи).

Каждый файл проверяется на правильность выделения информативной части ГК. При необходимости изменяется пороговое значение уровня сигнала, которое используется для выделения информативной части. Имена файлов с типичным произнесением ГК фиксируются. Затем происходит преобразование полученного множества файлов во множество файлов, содержащих только выделенную информативную часть ГК.

Затем определяются параметры кадра (25 мс – 200 отсчетов с частотой дискретизации 8 кГц) РС из заданного перечня, соответствующие информационным частям ГК. При необходимости создаются текстовые файлы, содержащие векторы параметров. Затем по известному файлу типичного произнесения ГК из соответствующей последовательности векторов параметров РС создается МСМП с начальными значениями числовых характеристик. Далее запускается итерационная процедура Баума – Уэлча и создается итоговая МСМП с конечными значениями числовых характеристик.

После создания всех МСМП для заданного перечня ГК проводится их проверка – проходит этап тестирующего распознавания ГК. Для этого используется еще одно - тестирующее - множество файлов информативных частей ГК. Результаты распознавания отражаются в окне статистики. Неверно и «неуверенно» (вероятности соответствия разным эталонам близки) распознанные файлы помечаются и в дальнейшем используются для анализа причин ошибки распознавания.

При анализе для каждого файла определяется (по алгоритму Витерби) наиболее вероятная последовательность состояний каждой МСМП из перечня всех эталонов распознаваемых ГК. Вычисляются также вероятности этих последовательностей.

С целью последующего выбора оптимальной МСМП из полученного их набора каждый вариант эталона ГК хранится в текстовом файле, который может редактироваться стандартными средствами операционной системы *Windows* 95/98.

Порядок выполнения работы

1. Получить задание с названиями двух голосовых команд и перечнем используемых для распознавания параметров речевого сигнала.

2. Сформировать обучающее и тестирующее множества звуковых файлов для первой голосовой команды (ГК1), контролируя правильность выделения информативной части. Из обучающего множества выбрать звуковой файл с типичным произнесением ГК1.

3. Сформировать обучающее и тестирующее множества файлов, содержащих только информативную часть ГК1.

4. Повторить пункты 2, 3 для ГК2.

5. Создать эталон (МСМП) ГК1.

6. Повторить пункт 5 для ГК2.

7. Провести распознавание ГК1, фиксируя ошибки распознавания и факты «неуверенного» распознавания. Пометить ошибочно и «неуверенно» распознанные файлы.

8. Провести анализ ошибок распознавания ГК1:

- просмотреть осциллограмму каждого ошибочно и «неуверенно» распознанного файла с результатом выделения информативной части ГК;
- вычислить вероятности *P*(ГК1, модель 1) и *P*(ГК1, модель 2), фиксируя изменение вероятностей при переходе от состояния к состоянию;
- дать заключение о причинах ошибки.

9. Повторить пункты 7, 8 для ГК2;

Содержание отчета

1. Структурная схема обработки речевых сигналов при выполнении лабораторной работы.

2. «Осциллограммы» и имена файлов типичных произнесений голосовых команд с помеченными информативными частями. Данные о пороговых значениях уровня сигнала.

3. Распечатки файлов начальных и конечных моделей голосовых команд. Данные о количестве файлов в обучающих множествах.

4. Данные о результатах распознавания команд с использованием тестирующих множеств файлов: размеры множеств, процент ошибки, перечень имен ошибочно и «неуверенно» распознанных файлов.

5. Анализ причин появления ошибок распознавания.

Контрольные вопросы

1. Каким образом в МСМП учитывается возможное увеличение длительности произнесения команды (по сравнению с типичным произнесением)?

2. Каким образом в МСМП учитывается возможное изменение параметров речевого сигнала команды (по сравнению с типичным произнесением)?

3. Каковы основные причины возникновения ошибок распознавания команд?

4. Каким образом изменение количества состояний МСМП влияет на точность распознавания команд?

5. Какие параметры МСМП определяют объем вычислительных затрат при распознавании?

Рекомендательный библиографический список

1. Рабинер Л.Р. Скрытые марковские модели и их применение в избранных приложениях при распознавании речи: Обзор // ТИИЭР. - М., 1989. - Т.77. - № 2. - С. 86 – 120.

2. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов /Пер. с англ.; Под ред. М.В. Назарова, Ю.Н. Прохорова. - М.: Радио и связь, 1981. - 496 с.

3. Елинек Ф. Распознавание непрерывной речи статистическими методами // ТИИЭР. – М.,1976. - Т.64. - № 4. – С. 131 – 160.

4. Ту Дж., Гонсалес Р. Принципы распознавания образов.– М.: Мир, 1978. - 416 с.

Оглавление

ЭФФЕКТИВНОЕ КОДИРОВАНИЕ И РАСПОЗНАВАНИЕ РЕЧЕВЫХ СИГНАЛОВ

Методические указания к лабораторным работам

Составитель
ЛЕВИН Евгений Калманович

Ответственный за выпуск – зав. кафедрой профессор О.Р. Никитин
Редактор Е.А. Амирсейидова
Корректор В.В. Гурова
Дизайн обложки К.Е. Левин

ЛР № 020275. Подписано в печать 26.04.02.
Формат 60x84/16. Бумага для множит. техники. Гарнитура Таймс.
Печать офсетная. Усл. печ. л.3,02. Уч.-изд. л. 3,22. Тираж 100 экз.
Заказ
Редакционно-издательский комплекс
Владимирского государственного университета.

600000, Владимир, ул. Горького, 87.