ИЗВЕСТИЯ
ТОМСКОГО ОРДЕНА ОКТЯБРЬСКОЙ РЕВОЛЮЦИИ И ОРДЕНА ТРУДОВОГО КРАСНОГО ЗНАМЕНИ ПОЛИТЕХНИЧЕСКОГО ИНСТИТУТА
имени С. М. КИРОВА

Том 194 1972

К ВОПРОСУ ТЕХНИЧЕСКОЙ РЕАЛИЗАЦИИ ИДЕЙ АНАЛИЗА РЕЧИ

Б. Н. ЕПИФАНЦЕВ

(Представлена научным семинаром кафедры вычислительной техники)

Рассмотрение работ по автоматическому распознаванию речевых сигналов [1] дает основание утверждать, что несмотря на значительные усилия, предпринимаемые в последнее время в этом направлении, имеющиеся результаты не позволяют удовлетворить запросы многочисленных технических приложений. Причин здесь много и, бесспорно, одна из них состоит в использовании большинством исследователей в своих экспериментах узкоспециализированной аналоговой аппаратуры с ее ограниченными возможностями.

В 1964 г. бесперспективность применения специализированных устройств для анализа речи была обоснована [2]. Взамен же аналоговой аппаратуры на всех этапах работы предлагалось использовать универсальные вычислительные машины [2]. Эта точка зрения до сих пор не претерпела существенного изменения, хотя авторы ряда работ [3, 4] считают целесообразным применять универсальные вычислительные машины на последнем этапе анализа — обработке признаков, выделяемых, в свою очередь, с помощью специализированных установок.

Указанные предложения по техническому воплощению идей анализа речи имеют определенную общность — они ничем не обоснованы. Как будет видно из дальнейшего, необходимость последнего вполне очевидна.

При вводе в ЭВМ и обработке специальной информации на ЭВМ приходится сталкиваться с рядом достаточно общих условий:

1. $V_c \leqslant V_{озу}$, т. е. объем сигнала должен быть меньше объема оперативного запоминающего устройства вычислительной машины.

2. Пропускная способность информационных каналов машины должна быть согласована с плотностью поступающей в ЭВМ информации.

3. Конструкция ЭВМ не должна подвергаться какого-либо рода переделкам, затрудняющим решение других задач.

4. Должен быть предусмотрен контроль соответствия введенного сигнала исходному.

5. При выборе схемы технической реализации идей анализа речи должен учитываться фактор стоимости машинного времени.

Действительно, классическая схема ввода специальной информации в электронную вычислительную машину (датчик—преобразователь аналог — код — ЭВМ) предлагает измерение и ввод каждого отсчета исходной функции, т. е. вводимый в ЭВМ объем информации есть так

называемое абсолютное описание входного сигнала [5], представляющее исследуемый сигнал со всей полнотой и точностью, определяемыми способами наблюдениями и разрешающей способностью средств наблюдения. При принятой для речевых сигналов частоте квантования $f_{кв} = 22$ *кгц* и динамическом диапазоне 45—48 *дб* [6] количество информации, вводимое в ЭВМ для описания слова длительностью, скажем, 1 *сек*, составляет около $200 \cdot 10^3$ *бит*. Эта величина превышает объем оперативного запоминающего устройства большинства универсальных вычислительных машин. Поэтому при использовании классической схемы ввода специнформации длительность исследуемой функции приходится сокращать до неоправданно малых величин. Например, для машины БЭСМ-2М, имеющей объем оперативного запоминающего устройства $V_{озу} \approx 80 \cdot 10^3$ *бит*, отрезок времени существования функции в лучшем случае ограничивается величиной 0,4 *сек*, что явно недостаточно, если учесть наличие корреляции в речи на более крупных участках, нежели 0,4 *сек* [7].

Далее огромная плотность информации (200 000 *бит/сек*), поступающей с преобразователя аналог—код, исключает все возможные способы введения спецсигналов в ЭВМ, основанные на последовательных принципах записи, так как последовательные принципы записи у современных машин допускают плотность поступления информации, не превосходящую $50 \cdot 10^3$ *бит/сек*. Применение же параллельных принципов записи, во-первых, ограничивает класс машин, пригодных для анализов речи, во-вторых, значительно усложняет структуру преобразователей аналог—код [8], доводя ее до необоснованной сложности при попытках ввести контроль соответствия введенного сигнала исходному [8].

Наконец, обработка абсолютного описания сигнала чрезвычайно неэкономична с точки зрения затрат машинного времени. Так, если взять рассмотренный выше сигнал, абсолютное описание которого составляет $200 \cdot 10^3$ *бит*, и положить, что машина оперирует 9 разрядными двоичными числами, то для вычисления, скажем, K-й точки функции автокорреляции данного сигнала

$$B(\tau_к) = \frac{1}{m} \sum_{i=1}^{m} f_i \cdot f_{i+к}$$

потребуется выполнить $22 \cdot 10_3$ умножений, $22 \cdot 10_3$—сложений и одно деление. Для вычисления же всех возможных точек приведенные цифры возрастают до $484 \cdot 10_6$ умножений, $484 \cdot 10_6$ — сложений и $22 \cdot 10_3$ — делений.

Если теперь учесть, что качество решения задач, соответствующих третьему и четвертому пунктам, находится в прямой зависимости от объема сигнала и что реализации одного и того же звука, произнесенные одним и тем же диктором, значительно отличаются друг от друга, попытки представления сигнала с высокой точностью [6] не имеют смысла.

Рассмотрим другую идею применения электронно-вычислительных машин для исследования речевых сигналов. Смысл ее состоит в том, что при помощи аналоговой аппаратуры из речевого сигнала выделяется какой-либо признак, который затем через преобразователь аналог—код водится в ЭВМ (смотрите, например, [3, 8]). Подобный способ позволяет, вообще говоря, «безболезненно» выделять интересующие признаки, используя для последующей их обработки универсальные машины.

Но что он дает?

До сих пор о способе кодирования смысловой информации в речи нет никаких данных. Опыт предыдущих исследований убедительно говорит о том, что надежды решить задачу распознавания речевых сиг-

налов, используя ряд известных интегральных преобразований, не имеют под собой почвы. Более того, имеющаяся предпосылка принципиальной возможности перехода от абсолютного описания сигналов к системе признаков не очевидна. Как известно [8], согласно этой предпосылке орган слуха осуществляет какое-то преобразование, в результате которого плотность информации о сигнале, поступающем в нервную систему, не превосходит 50 *бит/сек* и, следовательно, коэффициент компрессии, осуществляемый ухом, достигает величины 400—1000. Но последняя цифра будет соответствовать истине лишь тогда, когда при определении объема сигнала на входе будем опираться на те же понятия, на которые опирались психологи при определенном объеме сигнала на выходе, чего, как раз, в действительности не наблюдается.

Возвращаясь теперь к вопросу целесообразности исследования параметров на ЦВМ, следует сказать, что пока предельная величина коэффициента компрессии не будет строго установлена, считать возможным решение задачи распознавания речевых сигналов на уровне отдельных заранее выбираемых признаков нельзя. Поэтому использование специализированной аналоговой аппаратуры ограничивает последующие возможности рассматриваемой схемы применения ЭЦВМ. Другим недостатком, присущим этой схеме, является то, что контроль соответствия введенного сигнала исходному осуществляется путем трудоемкого процесса записи на шлейфовый осциллограф сигналов, которые вводятся в ЭЦВМ.

В поисках схемы исследования, которая бы удовлетворяла пунктам 1—5 и в то же время позволяла оперировать с сигналами, относительно которых имелась уверенность, что количество информации, содержа-

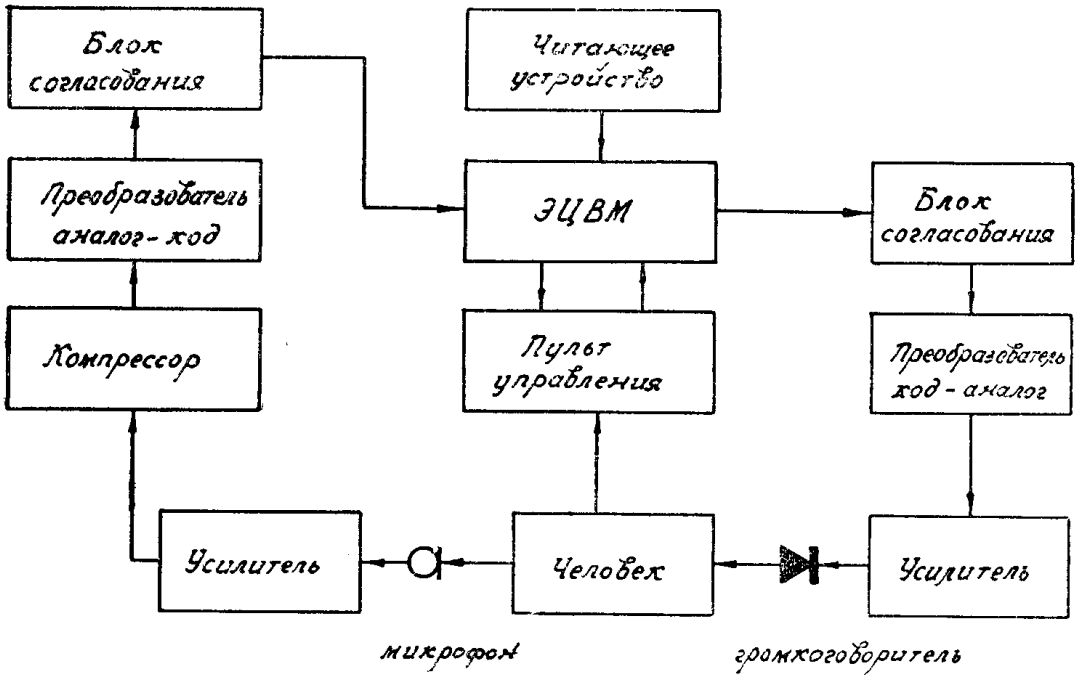

Рис. 1

щееся в них, достаточно для понимания сказанного, вспомним тот факт, что разборчивость «вокодерной» речи достаточна для понимания слов, а разборчивость клиппированного сигнала мало отличается от оригинального. В то же время объем «вокодерного» или клиппированного

сигнала более чем на порядок меньше неискаженного. Естественно напрашивается мысль использовать эти факты. Тогда схема исследования речевых сигналов будет выглядеть так (рис. 1).

Речевой сигнал предварительно подвергается компрессии, причем коэффициент сигнала должен быть таким, чтобы имелась возможность восстановления исходного сигнала с точки зрения его разборчивости. Через преобразователь аналог—код компрессированный сигнал вводится в ЭЦВМ. При коэффициенте компрессии $K_н = 10 \div 20$ условия 1, 2, 5 нарушаться не будут. Контроль соответствия введенного сигнала исходному в этом случае легко осуществить путем вывода из оперативной памяти введенного сигнала на синтезирующее устройство и его последующего прослушивания. Экспериментатор на слух легко может определить, была ли запись сигнала в ЭВМ удовлетворительной, либо ее следует стереть и записать новую.

Расмотренная схема имеет еще одно положительное качество.

Путем воздействия программными методами на записанный сигнал мы можем оценить различные преобразования над сигналом на слух, которые с помощью аналоговой аппаратуры поставить практически невозможно.

В заключение заметим, что рассмотренные схемы реализации идей анализа речи апробировались. В качестве ЭВМ использовалась БЭСМ-2М. Наиболее удачным оказалось применение для анализа киппированной речи. Отрезок сигнала, который было можно записать в МОЗУ, превышал 7 *сек.*

ЛИТЕРАТУРА

1. Н. Г. Загоруйко. Г. Я. Волошин, В. Н. Елкина. Автоматическое опознавание звуковых образов (обзор литературы). Вычислительные системы, вып. 14, 1964.

2. Н. Г. Загоруйко. Об обмене устной информации между человеком и вычислительными системами. Вычислительные системы, вып. 10, 1964.

3. О. А. Петров. Статистическая обработка первичных признаков речевых сигналов с использованием электронно-вычислительной машины (ЭВМ). «Опознавание образов. Теория передачи информации», 1965.

4. В. П. Трунин-Донской, А. С. Фирер. Ввод в ЭВМ БЭСМ-2 некоторых речевых признаков. «Ввод и обработка специальной информации в ЭВМ», вып. 4 М., ВЦ АН СССР, 1964.

5. А. А. Харкевич. О выборе признаков при машинном опознавании. Изв. АН СССР, ОТН, техническая кибернетика, № 2, 1963.

6. Г. Я. Волошин. Преобразователь аналог—цифра для ввода речевых сигналов в ЭВМ. Вычислительные системы, вып. 10, 1964.

7. А. Н. Величкин. Статистическое исследование речевого процесса. «Электросвязь», VIII, № 8, 1961.

8. Моль. Теория информации и эстетическое восприятие. «Мир», 1966.