ЭМПИРИЧЕСКОЕ ПРЕДСКАЗАНИЕ И РАСПОЗНАВАНИЕ ОБРАЗОВ
(Вычислительные системы)
1976 год Выпуск 67

УДК 681.142.37

ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА НЕКОТОРЫХ ОСОБЕННОСТЕЙ РЕЧЕВОГО ВВОДА ДАННЫХ

С.В.Голубцов, В.А.Буртасов, В.А.Данилов

Вопрос о целесообразности ввода информации в форме устной речи обсуждается в течение уже ряда лет [I]. Сторонники речевого ввода обосновывают преимущества последнего возможностью повышения скорости ввода информации, повышением пропускной способности и комфорта работы оператора, освобождением его рук и зрения для выполнения других функций. Возражения против применения речевого ввода основываются на том, что существующие методы и алгоритмы распознавания имеют серьёзные ограничения, связанные с объемом словаря, достоверностью распознавания, чувствительностью к смене дикторов и внешним акустическим условиям. Допускается, что эти ограничения могут свести на нет ожидаемый выигрыш от применения речевого ввода. Например, вследствие ограниченной достоверности, вероятно, нельзя отказаться от визуального контроля вводимой информации, и зрение оператора остается по-прежнему занятым. Ставится под сомнение повышение степени комфорта работы оператора, поскольку требование отсутствия оговорок при произнесении команд будет держать его в постоянном напряжении и быстро утомлять.

Ту или иную точку зрения в вопросе о применимости речевого ввода информации, по-видимому, невозможно доказать теоретически. Необходимо экспериментальное изучение устройств речевого ввода, работающих в составе реальных управляемых комплексов, и сопоставление условий работы оператора, использующего речевой ввод, с другими известными способами управления. Для экспериментального изучения особенностей речевого ввода было раз-

работано автономное устройство, названное автоматом К-2,позволяющее после настройки на рабочий словарь подключаться к различным объектам управления.

О с н о в н ы е х а р а к т е р и с т и к и а в т о м а т а К-2. При разработке устройства был принят алгоритм (описанный в [2]),обеспечивающий при моделировании его на ЭЦВМ распознавание нескольких десятков слов, произносившихся различными дикторами, с достоверностью выше 90%. Слова произносились диктором в шумостойкий микрофон типа ДЭМШ-Iа, укрепленный на держателе, смонтированном на оголовье для телефонных наушников. Время выдачи ответа автоматом не превышает 0,4 сек после окончания произнесения слова. Слова при произнесении должны разделяться паузами длительностью не менее 0,5 сек. Результаты распознавания высвечиваются на табло. В автомате не предусмотрен отказ от решения (ответ "не знаю"), за исключением случая очень тихого произнесения, когда уровень речи не превышает порога отсечки паузы. Кроме распознавания слов ,в автомате К-2 предусмотрен режим распознавания команд.Команды могут состоять из одного или двух слов, входящих в словарь устройства. Общее количество команд равно 90, в том числе двусловных - 58.Для повышения достоверности распознавания команд в алгоритм была введена операция коррекции. Коррекция состояла в замене одного из слов двусловной команды на наиболее вероятный заменитель в случае появления невозможного сочетания слов. Вероятные заменители определялись заранее по статистике ошибок распознавания слов.

После изготовления и настройки автомат К-2 был подвергнут лабораторным испытаниям, основные результаты которых излагаются ниже.

Р е з у л ь т а т ы л а б о р а т о р н ы х р е ч е в ы х и с п ы т а н и й а в т о м а т а К-2. При проведении лабораторных испытаний автомата К-2 были поставлены следующие задачи:

а) определение достоверности распознавания слов на голосах представительной группы дикторов;

б) определение достоверности распознавания команд;

в) оценка способности операторов к обучению;

г) изучение внешних факторов, влияющих на условия распознавания, в частности:

- уровня акустических шумов в помещении;
- положения микрофона в гарнитуре;
- уровня входного речевого сигнала;
- вариативности произнесения слов диктором.

Для определения достоверности распознавания слов было привлечено 25 дикторов (20 мужских и 5 женских голосов). Около половины дикторов не имело опыта чтения перед микрофоном,а опытом работы с автоматом К-2 обладало только 6 человек (трое из них входило в число непосредственных разработчиков устройства, остальные эпизодически привлекались как дикторы в процессе настройки автомата). У некоторых из дикторов имелись заметные на слух дефекты речи. Каждому диктору перед проведением зачетного эксперимента давалась возможность в течение 5-7 мин освоиться с микрофоном, подобрать его положение в гарнитуре и потренироваться путем произнесения слов, входящих в состав словаря,с контролем результата распознавания по световому табло. В некоторых случаях при этом производилась подстройка уровня речи по индикатору, входящему в комплект автомата ,путем изменения положения движка потенциометра входного усилителя. Затем диктору предлагалось произнести по 3 раза каждое из 62 слов. Результаты распознавания слов, высвечиваемые на табло, фиксировались сотрудником, проводящим эксперимент. По результатам распознавания составлялись матрицы; ниже приводятся основные итоговые данные.

Достоверность распознавания по дикторам значительно колеблется от 98,4% до 36,7%. На рис.I приведена гистограмма значений достоверности η (кривая I). Из нее следует, что вся группа может быть разбита на 4 класса - "отличных" ($100\% \geq \eta > 95\%$), "хороших" ($95\% \geq \eta > 85\%$),"посредственных" ($85\% \geq \eta > 60\%$) и "плохих" ($\eta \leq 60\%$) дикторов. Количественно из всей группы 25 дикторов "отличных" оказалось 8%, "хороших" - 32%, "посредственных" - 52% и "плохих" - 8%.

Сопоставление приведенной классификации с конкретными дикторами показывает, что основное влияние на результат распознавания оказывает способность диктора четко, без редукции произносить звуки в словах (но не диктовать), а также сохранять

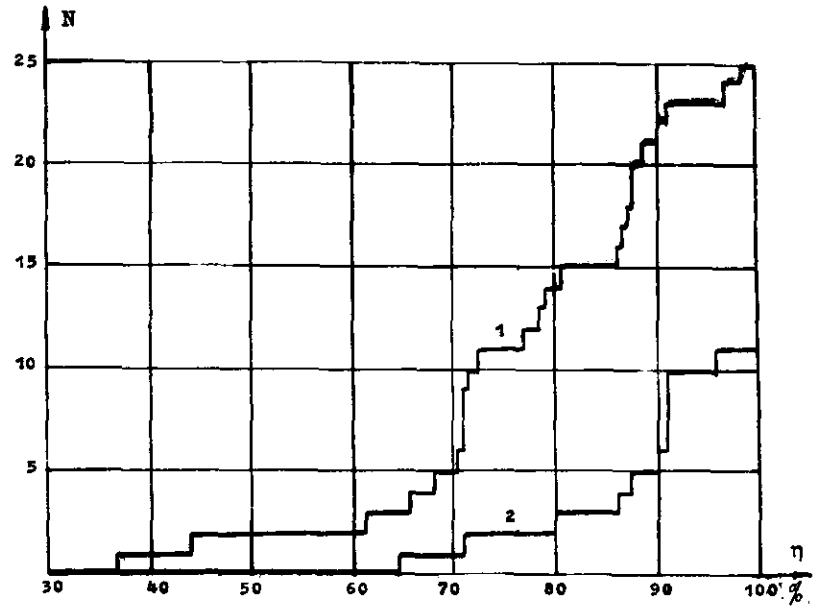

Рис. I

темп и уровень речи постоянным в течение всего сеанса.Одновременно было отмечено, что отдельные дефекты речи (назализация, присвистывание) часто не оказывают существенного влияния на результаты.

Средняя достоверность распознавания по всей группе 25 дикторов составляет 76,8%. При оценке достоверности по пяти лучшим дикторам достоверность равна 92,8%, для десяти лучших голосов, составляющих группу "хороших" и "отличных" дикторов,достоверность распознавания имеет значение 89,8%. Мужские голоса распознаются с достоверностью 79,4%,женские - 66,8%.Это различие вполне объяснимо тем, что при обучении привлекались только мужские голоса.

Вследствие использования в качестве дикторов лиц, совершенно не подготовленных для этой цели, и при отсутствии тренировки в необходимом объеме полученные данные содержат много

случайных результатов. Для их устранения средние значения достоверности по трем прочтениям 25 дикторов были обработаны согласно методике, изложенной в [3], и определено вероятное значение достоверности с надежностью $\delta = 0,9$.Оно оказалось равным $\eta_{вер.} = 79,9 + 3,7\%$.

Достоверность распознавания отдельных слов для группы из 25 дикторов колеблется от 97,4% до 42,7%, причем не нашлось ни одного безошибочно принятого слова. Для десяти лучших дикторов встретилось I2 слов (около 20% словаря), распознававшихся безошибочно.Следует отметить, что большинство слов у I0 дикторов имеет близкие значения достоверности и понижение на отдельных словах вызвано скорее всего недостаточной корректировкой их эталонов и при необходимости может быть устранено в дальнейшем.

Таблица

Слово	Достоверность,%	
	25 диктор.	I0 диктор.
ноль	74,7	90,0
один	82,7	93,3
два	74,7	83,3
три	76,0	96,7
четыре	76,0	I00,0
пять	76,0	90,0
шесть	74,7	90,0
семь	73,3	90,0
восемь	76,0	93,3
девять	68,0	73,3
	75,2	90,0

В состав словаря входили названия всех однозначных десятичных цифр. Поскольку цифры необходимы, как правило,в любом словаре, достоверность их распознавания приведена отдельно в таблице. В среднем по 25 дикторам достоверность распознавания цифр составила 75,2%, а по I0 диктора - 90%. Лучше других (безошибочно) у I0 дикторов распознается слово "четыре", хуже всех - "девять". Средняя достоверность распознавания цифр соответствует достоверности распознавания всего словаря.

Следующий этап испытаний состоял в оценке достоверности распознавания полного набора 90 команд. В отличие от предыдущих испытаний диктору после непродолжительной тренировки предлагалось прочитать список команд только один раз. Хотя в качестве дикторов были привлечены сотрудники, уже участвовавшие в испытаниях автомата в режиме распознавания слов, при чтении команд возникли дополнительные трудности,связанные с необходимостью выдержки паузы между словами двуслов-

ной команды.Многие дикторы стремились произносить слова команды если не слитно, то с минимальной паузой, недостаточной для завершения распознавания первого слова,занимающего 0,3–0,4 сек. Кроме того, заранее зная оба слова команды, они непроизвольно вносили в слова при их произнесении изменения, характерные для слитной речи. Поскольку эти факторы были выявлены уже в ходе испытаний, сколько-нибудь серьёзной тренировки дикторов для чтения двусловных команд организовать не удалось, и отмеченные особенности нашли свое отражение в приводимых ниже результатах.

В испытаниях приняли участие II дикторов (I0 мужских и I женский голос),четыре диктора прочитали команды дважды).В состав группы были включены дикторы, отнесенные по результатам распознавания слов к "отличным" и "хорошим". Значения достоверности по всей группе голосов колеблются в пределах от 64,5% до 95,6%.

На рис.I приведена гистограмма значений достоверности распознавания команд (кривая 2). Средняя достоверность распознавания команд по всем II дикторам составляет 86,2%. После отбрасывания случайных результатов по методике [3] вероятное значение достоверности оказывается равным $\eta_к$ = 9I,I±I,8%, δ = 0,9. При оценке по пяти лучшим голосам достоверность распознавания 90 команд составляет 92,9%. Одна из особенностей режима распознавания команд состоит в возможности появления ответа "не знаю". Ответ "не знаю" появляется в случае, когда было принято невозможное для автомата К-2 сочетание слов, которое не может быть исправлено на "разрешенное" сочетание имеющимися в схеме правилами коррекции. При ответе "не знаю" на табло автомата не зажигается никакой надписи. В среднем по II дикторам "не знаю" встречается приблизительно в 5% всех ответов.

Одним из средств повышения достоверности является согласование характеристик распознающего автомата и диктора. Характеристики дикторов по возможности были учтены в эталонах команд, заложенных в схеме устройства. Поэтому интересно оценить возможность использования другого пути– подстройки голоса оператором под особенности автомата.

В самой постановке вопроса нет,по-видимому, ничего необычного, поскольку известно, что для очень многих практических задач необходима тренировка операторов, подчас весьма длительная и трудоемкая.

Уже в первых опытах было замечено,что в ряде случаев можно добиться правильного срабатывания автомата,слегка растягивая гласные в слове, подчеркивая фрикативные или дрожащие. В то же время отмечалось, что достаточно жестких формализованных правил подстройки голоса выработать не удается. При проведении лабораторных испытаний автомата К-2 изучению возможности подстройки голоса оператора было уделено определенное внимание. Во-первых, было решено оценить роль визуального контроля. С этой целью вначале была определена средняя достоверность по каждому из трех произнесений 25 дикторов. Предполагалось, что благодаря визуальному контролю диктор при повторном произнесении исправляет свои ошибки (хотя при этих измерениях диктору подобной инструкции и не давалось).Это предположение не подтвердилось. На рис.2 приведена зависимость достоверности распознавания слов η для 25 дикторов от номера произнесения n (кривая 2). Оказалось, что кривая имеет тенденцию некрутого падения с ростом n . Подобные данные были получены и на трех опытных дикторах,участвовавших в следующем эксперименте. Этим диктором было предложено прочитать каждое слово еще по три раза, но не глядя на световое табло, благодаря чему дикторы не знали результатов распознавания. Полученные данные показывают, что достоверность распознавания для всех без исключения дикторов заметно упала (в среднем более чем на I0%). Эти данные прямо свидетельствуют о важности визуального контроля, с помощью которого диктор каким-то образом контролирует свою речь.

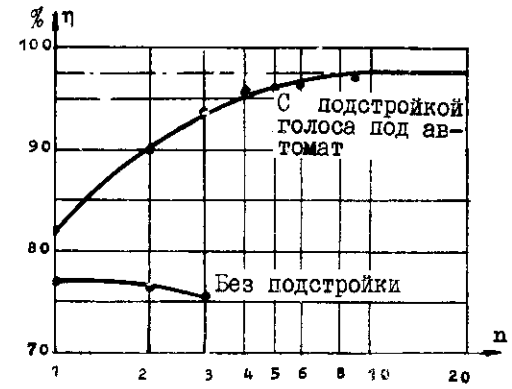

Рис. 2

Последний эксперимент в этом направлении состоял в определении возможности оператора добиться от автомата требуемой реакции. В опыте приняли участие 9 дикторов (7 мужских и 2 жен

ских голоса), которым было предложено произносить каждое слово до тех пор, пока автомат не даст правильного ответа или диктор почувствует, что правильного ответа добиться невозможно. При проведении опыта подсчитывалось количество попыток, потребовавшееся для получения правильного ответа, а также количество случаев отказа диктора от дальнейших попыток. При подсчете результатов оказалось, что из всех 558 слов-произнесений с первого раза был получен правильный ответ в 82,5% случаев.В среднем на одну команду каждый диктор затратил I,6 произнесения, причем I2 слов (2,9%), повторенные дикторами по I0-I5 раз, так и не были правильно распознаны.

Из всех 62 слов словаря у всех 9 дикторов с первого произнесения правильно были распознаны I4 слов; не было ни одного слова, потребовавшего повторения больше чем у четырех дикторов. Слов, потребовавших повторения у четырех из девяти дикторов, оказалось всего четыре. На I0 различных словах словаря дикторы отказались от дальнейших попыток, однако у каждого диктора это были свои слова; только на одном из слов отказы были сразу у двух дикторов. У четырех из 9 дикторов вообще не было ни одного отказа.

Динамика процесса обучения иллюстрируется кривой I на рис.2. Кривая построена по средним значениям для всех дикторов, поскольку для отражения их индивидуальных особенностей было недостаточно материала. Из кривой следует, что путем многократного повторения с визуальным контролем результата можно добиться весьма высокой достоверности распознавания - почти 98%,причем по сравнению с первым произнесением достоверность повышается более чем на I5%. В то же время наиболее крутой участок кривой приходится на первые четыре произнесения,после чего исправляются только единичные ошибки. Сопоставление кривых I и 2 (рис.2)показывает важность установки для оператора: при получении данных кривой I, в отличие от кривой 2, была дана инструкция на подстройку голоса по результатам распознавания.

Таким образом, обучение оператора работе с распознающим автоматом является весьма важным этапом при организации речевого ввода информации. Это подтверждают и субъективные ощущения большинства дикторов, считающих, что они могут значительно повысить свои результаты, если им будет дана возможность предварительной тренировки в достаточном объеме.

Важной характеристикой автомата для ввода информации в форме устной речи является его способность работы в условиях шума. Для уменьшения внешних шумов было решено использовать шумостойкий микрофон типа ДЭМШ-Iа, что в значительной мере повысило невосприимчивость устройства к посторонним звукам.Для автомата К-2 это особенно важно, поскольку в его алгоритме не предусмотрено практически никакой защиты от посторонних слов, не входящих в словарь: если какой-либо звук в схеме автомата не классифицируется как пауза, на выходе устройства обязательно появится ответ в виде одного из слов словаря,наиболее близкого по своим акустическим характеристикам к принятому звуку. При проведении испытаний не принималось никаких мер по уменьшению уровня шума в помещении, где одновременно находилось и работало с различной аппаратурой I0-I5 человек. При проведении испытаний уровень шумов помещения постоянно фиксировался шумомером; шум колебался в пределах от 45 до 65 дб; в среднем по всем основным измерениям он составлял 57 дб. Шум возникал от работы вентиляторов, двигателя шлейфного осциллографа, а также от переговоров сотрудников, работающих в комнате. Все эти шумы почти не оказывали влияния на результаты. Искажения вызывали, как правило, помехи импульсного характера: щелчки от включения электромагнитов магнитофона, удары молотка, хлопанье дверью и др. Импульсные помехи в зависимости от момента их появления или индицировались как одно из слов, или искажали произносимое слово (если помеха появлялась непосредственно перед началом произнесения слова, то присоединялась к нему). Было замечено также, что источником помех часто является сам диктор, непроизвольно создающий посторонние звуки перед самым началом произносимого слова (щелчок языком, придыхание и др.). Помехи этого рода удавалось, как правило, устранить в процессе тренировки.

Для определения шумостойкости К-2 были поставлены два эксперимента. В задачу первого из них входило получение зависимости достоверности распознавания слов от уровня внешнего шума $P_ш$, создаваемого искусственно.

Для создания шума был использован магнитофон МАГ-8м,который воспроизводил записанный на кольце ленты шум со спектром Хотта. Магнитофон располагался на расстоянии I,5-2 м напротив диктора, уровень шума устанавливался путем регулировки

усилителя магнитофона и контролировался по шумомеру. При проведении эксперимента было обнаружено, что в условиях шумов порядка 80 дб в автомате нарушается нормальная работа схемы отделения сигнала от паузы, в связи с чем был поднят приблизительно на 20% порог выделения паузы и было решено весь эксперимент в шумах проводить с таким порогом.

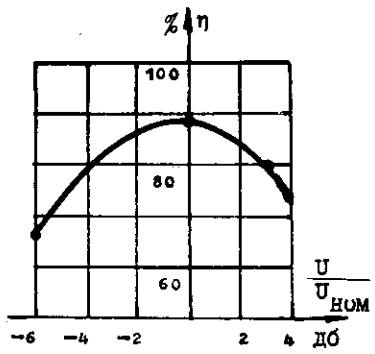

Рис. 3

Результаты измерения достоверности в условиях шумов, проведенные на голосах трех дикторов, читавших по три раза каждое слово, приведены на рис.3. Сплошная линия соответствует усредненным значениям достоверности по дикторам, а заштрихованная область определяет разброс отдельных измерений. Кривая η ($P_ш$) оказалась немонотонной, имеющей максимум в области $P_ш$ = 60 дб. По-видимому, это обусловлено, с одной стороны, влиянием повышенного порога выделения паузы, при котором шум умеренного уровня как бы "помогает" отделять слово от паузы без потери слабых звуков. С другой стороны, в условиях шумов диктор непроизвольно форсирует голос, что приводит к более четкой артикуляции. Из рис.3 следует, что автомат "К-2" сохраняет работоспособность в шумах с уровнем до 70-75 дб; при этом достоверность распознавания слов изменяется в пределах ±3%.

Для оценки реакции автомата на помехи речевого происхождения был поставлен еще один опыт. При идеальной настройке можно ожидать, что на речь произвольного содержания автомат должен распределять ответы с равной вероятностью по всем словам словаря. В то же время при существовании приоритета для некоторых слов последние должны появляться в ответах автомата чаще.

В качестве "речи произвольного содержания", обладающей достаточной языковой представительностью, была использована серия из трех стандартных словесных артикуляционных таблиц (см. [3], табл. I-3). Двум дикторам было предложено прочитать указанные таблицы перед микрофоном К-2, а в качестве результата опыта фиксировалось количество случаев высвечивания каждого из 62 слов на световом табло (число случаев активизации выходов автомата).

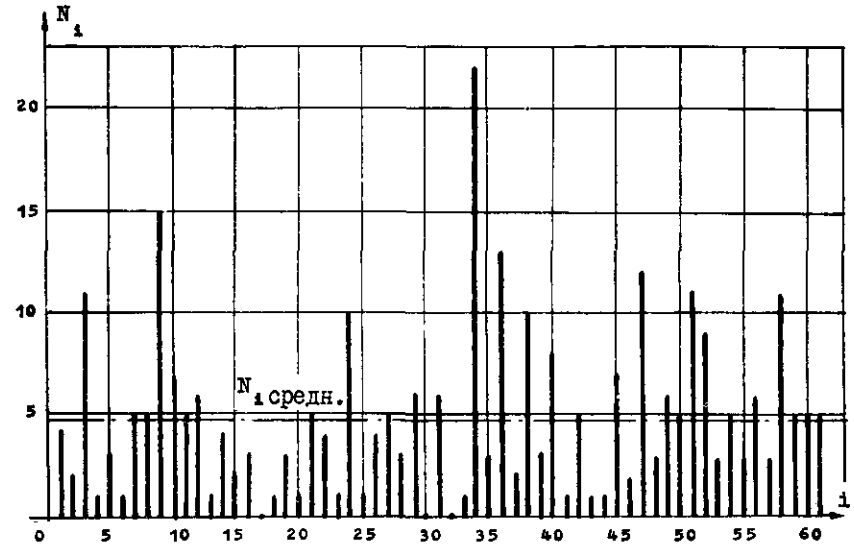

Рис. 4

Полученные данные, усредненные по обеим дикторам, приведены на рис.4, где по оси абсцисс отложены номера выходов автомата i, а по оси ординат – число N_i ответов для каждого выхода. Среднее число ответов на I выход в условиях опыта равнялось N_{icp} = 4,8. Из рисунка видно, что около половины выходов имеет число срабатываний, близкое к N_{icp}, однако ряд выходов не имел ни одного случая активизации, в то время как на некоторых из них число таких случаев значительно превышает среднее. Больше всего случаев активизации возникло на слове "автоматически". Данные рис.4 могут быть использованы при корректировке эталонов слов.

При проведении испытаний была сделана оценка стабильности результатов измерений и степени влияния некоторых дестабилизирующих факторов. Во-первых, был определен разброс достоверности распознавания для различных произнесений слов и команд.

Оценка проводилась по среднелинейному отклонению σ, определяемому как

$$\sigma = \frac{1}{n} \sum_{j=1}^{n} (\eta_j - \bar{\eta}),$$

где n – общее число измерений, η_j – значение достоверности для единичного измерения, $\bar{\eta}$ – среднее по всем η_j.

Среднелинейное отклонение достоверности одного произнесения для пяти лучших дикторов, читавших слова по три раза, равно 2,56%. Аналогичная величина, определенная для случая, когда дикторы читали слова в различных сеансах, оказалась равной 2,13%, т.е. примерно равной случаю чтения слов подряд. Среднелинейное отклонение достоверности распознавания команд, определенное при повторном их произнесении 4 дикторами, оказалось равным 2,9%.

При проведении измерений было замечено, что результаты распознавания зависят от положения микрофона относительно рта говорящего, в связи с чем была сделана количественная оценка этого фактора. Эксперимент проводился с одним диктором. В качестве эталонного было принято положение микрофона напротив правого угла рта, плоскость микрофона параллельна плоскости щеки, расстояние до микрофона – 15 мм. Была замерена достоверность распознавания слов при уменьшении расстояния от микрофона до угла рта до 10 мм и увеличении его до 40 мм, при перемещении микрофона вперед и назад на 30 мм и при повороте плоскости микрофона относительно плоскости щеки на $+30^0$ (по часовой стрелке) и -30^0. Было найдено, что наиболее сильно (на 10–15%) снижается достоверность при повороте микрофона против часовой стрелки и сдвиге его назад. Последнее вполне объяснимо, поскольку начинает сказываться экранировка звука щекой диктора.

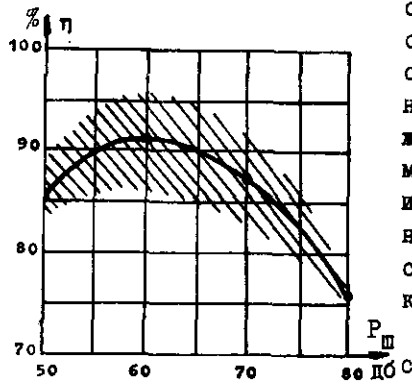

Рис. 5

В автомате К-2 отсутствует система АРУ, и в то же время ряд узлов, например схема отделения сигнала от паузы, построен на пороговых критериях. В связи с этим

при работе с устройством необходимо поддерживать некоторый оптимальный уровень речи. Практически это осуществляется по прибору, реагирующему на средний уровень речи, и в задачу диктора при тренировке входит подобрать громкость произнесения, ориентируясь на прибор. Для оценки критичности установки уровня была определена зависимость достоверности распознавания слов, произносившихся с нормальной громкостью при преднамеренном изменении коэффициента усиления входного усилителя на $^{+3}_{-6}$ дб. Полученная кривая приведена на рис.5. Из нее следует, что при изменении уровня речи относительно номинального значения достоверность понижается с крутизной приблизительно 3%/дб.

Из вышеизложенного можно сделать выводы:

1. Применение сравнительно простого двухэтапного алгоритма распознавания, предусматривающего использование спектрально-временного описания речи на первом этапе и линейного нормирования по времени на втором, обеспечивает достоверность распознавания слов около 90% при условии предварительного отбора дикторов и 80% в среднем для произвольного диктора.

2. Важное значение имеют тренировка и способность подстройки голоса диктора под особенности устройства, позволяющие значительно повысить достоверность распознавания слов.

3. К эксплуатационным недостаткам автомата К-2, выявившимся в ходе испытаний, относятся: а) необходимость пословного произнесения команды; б) сложность и неоперативность смены словаря; в) отсутствие автоматической регулировки уровня речи.

4. Несмотря на имеющиеся недостатки, автомат К-2 может быть использован для экспериментального изучения особенностей ввода информации на реальных объектах управления.

Литература

1. ГОЛУБЦОВ С.В. Задачи и перспективы распознавания речи. –В кн.: Труды УІ Всесоюзн. семинара АРСО-ІУ, Таллин, 1972, с. 64–73.

2. ОСАДЧИЙ Ю.Н. Оценка возможности распознавания ограниченного набора команд с использованием субфонемных последовательностей. –В кн.: Труды Акуст.ин-та. Вып.12. М., 1970, с.60–63.

3. ПОКРОВСКИЙ Н.Б. Расчет и измерение разборчивости речи. М., "Связьиздат", 1962.

Поступила в ред.-изд.отд.
7 мая 1975 года