EESTI NSV TEADUSTE AKADEEMIA TOIMETISED. XV KÖIDE
FÜÜSIKA-MATEMAATIKA- JA TEHNIKATEADUSTE SEERIA. 1966, NR. 3

ИЗВЕСТИЯ АКАДЕМИИ НАУК ЭСТОНСКОЙ ССР. ТОМ XV
СЕРИЯ ФИЗИКО-МАТЕМАТИЧЕСКИХ И ТЕХНИЧЕСКИХ НАУК. 1966, № 3

Э. КЮННАП

УСТНЫЕ КОМАНДЫ В СИСТЕМАХ УПРАВЛЕНИЯ

Обзор

Проблема автоматического распознавания звуков речи сравнительно молода. Повышение интереса к ней стало особенно заметно после второй мировой войны в связи с увеличением нагрузки каналов связи и поисками путей их уплотнения. Ученые США, Швеции, Англии, Японии и других стран усиленно работали в этой области. За 10 лет (1950—1960) только в восьми специальных журналах было опубликовано 275 статей по автоматическому распознаванию и синтезу речевых сигналов [158]. В последние годы издан целый ряд работ советских [11, 12, 29] и зарубежных авторов, посвященных вопросам состояния исследования речи [54, 119], а также работы, анализирующие отношения между машиной и человеком [114, 162].

В настоящем обзоре систематизированы некоторые данные, полученные в итоге исследования речи, и практические применения этих результатов.

1. Уплотнение канала сзязи

Как известно, речевой сигнал состоит из суммы отдельных колебаний разной частоты и амплитуды. При разложении его в ряд суммирование может быть осуществлено либо по элементам, равноотстоящим по частоте (ряд Фурье), либо по элементам, равноотстоящим во времени (теорема Котельникова). В первом случае речевой сигнал подвергается гармоническому анализу и по каналу передаются амплитуды (и фазы) гармоник; в приемном пункте речевой сигнал восстанавливается по этим спектральным коэффициентам, установленным анализатором в начальном пункте канала связи. Во втором случае по каналам связи через дискретные промежутки времени передаются импульсы, величина которых пропорциональна мгновенным значениям функции, взятым через интервалы Δt. На приемном конце эти импульсы проходят через фильтр, выход которого постоянно суммируется.

Передаваемая информация содержит наряду с полезными сведениями еще и шумы. В качестве измерителя, характеризующего уровень сигнала по сравнению с помехой, введена величина $H = \log p/p_{\text{п}}$, где p и $p_{\text{п}}$ — средние мощности сигнала и помехи соответственно. Произведение трех величин $V = TFH$ называется объемом сигнала. Канал связи характеризуется также тремя величинами: $T_{\text{к}}$ — промежуток времени, в течение которого канал включен, $F_{\text{к}}$ — полоса частот, пропускаемая каналом, и $H_{\text{к}}$ — уровни мощностей аппаратуры канала. Произведение $V_{\text{к}} = T_{\text{к}} F_{\text{к}} H_{\text{к}}$ называется емкостью канала. Для обеспечения пропускания сигнала через канал должно быть выполнено условие $V_{\text{к}} \geqslant V$.

Уплотнение канала связи может быть осуществлено при помощи деформации одной из пар этих величин при неизменной третьей. Деформация сигнала производится сжатием его на передающем конце и соответствующим расширением на приемном. Изменение H, T и F соответствует изменению усиления, задержке сигнала при помощи

линии задержки и частоты соответственно. Например, если записать речевой сигнал на магнитную ленту, передать его в два раза быстрее, чем он был записан, а в приемном пункте проиграть в два раза медленнее, то объем передаваемой информации останется без изменений, а передача будет произведена в два раза быстрее при удвоенной частоте.

Компандирование речевого сигнала, т. е. его компрессия в передающем и экспандирование в приемном конце передачи, может быть частотным, амплитудным и временным. Одним из крайних видов амплитудного компандирования речевого сигнала является клиппирование. В этом случае амплитуда речевого сигнала ограничивается двумя уровнями и передаются только точки, в которых функция меняет знак.

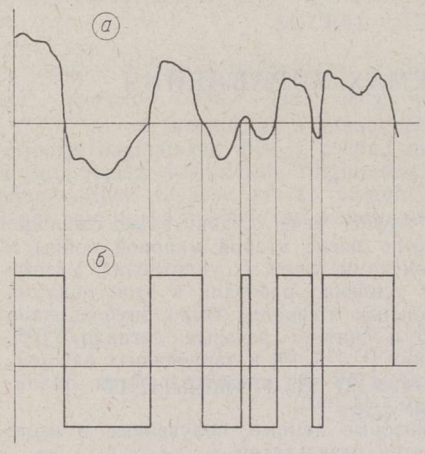

Рис. 1. Клиппирование речевого сигнала: *а* — оригинальный речевой сигнал; *б* — клиппированный речевой сигнал.

Известно, что чем меньше основание системы счисления, тем больше требуется разрядов для изображения одних и тех же чисел. Оптимальной системой была бы система с основанием числа *е*. Однако такую систему практически осуществить невозможно и за основание приходится принимать либо 2, либо 3. Наибольшее распространение получила двоичная система, хотя троичная была бы рациональнее, так как три ближе к значению *е*, чем два. Клиппирование речи (рис. 1) соответствует двоичной системе исчисления. Как уже заметил И. Ликлайдер [111, 112], при ограничении речевого сигнала двумя уровнями в сигнале остается еще достаточное количество сведений, чтобы разборчивость сохранилась на требуемом уровне. Технические же условия импульсного телефона предусматривают 128 уровней. Следовательно, клиппирование речевого сигнала уменьшает его объем в семь раз.

Разборчивость речи увеличивается, если речевой сигнал перед клиппированием дифференцируется. При этом частота клиппированного сигнала увеличивается и передается расположение экстремальных точек оригинального сигнала. Частичное компандирование осуществляется путем деления частоты сигнала на передающем конце и соответствующего умножения ее на приемном конце [36, 156]. Сужение спектра получается только до шести раз с большими искажениями речи, и поэтому такое компандирование особых перспектив не имеет. Но сокращение частного диапазона речи путем ограничения его сверху и снизу дает обратный эффект. Человеческий голос охватывает диапазон частот от 50—60 *гц* до 15—20 *кгц*. Если от телефонного разговора требуется только разборчивость и узнаваемость собеседника по голосу, то верхний предел частоты может быть понижен до 2,5—3 *кгц*. Если же потребовать только разборчивости, то объем сигнала может быть еще уменьшен. В условиях шумов ограничение при телефонной передаче речи сверху частотой до 3500 *гц*, а снизу до 300 *гц* повышает разборчивость речи [31, 130, 136].

При временном компандировании из речи исключаются определенные временные интервалы, а возникающие паузы заполняются другой передачей. В приемном пункте паузы данной речи заполняются слушателем благодаря действию мозга [25, 26, 39, 73]. Однако достигаемая при этом низкая степень компандирования (до двух) и снижение разборчивости говорят о том, что и этот метод особых перспектив не имеет [103]. Лучшие результаты временного уплотнения канала связи дает использование пауз между словами и фразами естественной речи, а также пауз на одной линии, когда собеседник говорит по парной линии. Переключение отдельных разговоров по одной линии между паузами других разговоров производится клиппированием речевого сигнала. При такой системе можно уплотнить канал связи до 4—6 раз.

Параметрические методы компандирования, хотя и позволяют значительно больше уплотнить канал связи, но нарушают микроструктуру речевого сигнала: по каналам передаются только параметры, полученные анализатором речевого сигнала, а в приемном конце эти переданные параметры управляют синтезатором речи. Таким образом, параметрические методы компандирования связаны с автоматическим распознаванием и синтезированием речевых сигналов.

Устройства, передающие речевой сигнал параметрическим методом, получили название вокодеров [59, 60]. В полувокодерных устройствах параметрическим методом передается только верхняя часть речевого сигнала, а нижняя часть передается непосредственно [69]. Кроме уплотнения каналов связи, вокодеры могут быть использованы для засекречивания телефонных разговоров [2, 102, 157].

2. Исследование образования звука

Ряд работ посвящен исследованию функционирования уха [14, 84, 92, 114] и голосовых связок [64, 89, 94, 165]. В [117] исследовано шесть мужских голосов путем применения анализа Фурье одного периода колебания давления в речевом тракте, в [105, 106] определена зависимость основного тона от давления воздуха в голосовой щели, а в [24] исследовано звукообразование и определение формантов анатомическими измерениями речевого тракта с использованием рентгеновских лучей и обработкой данных на ЭВМ.

Звук создается импульсами голосовых связок, частота колебания которых определяет высоту основного тона. Эти импульсы звука имеют дискретный спектр с большим числом гармоник в широком диапазоне частот. Частота основного тона находится в среднем в пределах 80—350 *гц*. По данным одних авторов, амплитуды гармоник почти одинаковы в широком диапазоне частот [17], а по данным других — они с увеличением частоты равномерно уменьшаются [64]. Звуковой сигнал получается от звука связок при прохождении его через резонирующие полости рта и носа, вследствие чего амплитуды некоторых гармоник звука связок уменьшаются или совсем подавляются, другие — усиливаются, образуя резонансные пики (так наз. форманты), измерению которых также посвящен ряд работ [65, 66, 165].

Каждому звонкому звуку соответствует своя комбинация формантов. По данным [3], в русской речи гласные *у*, *о*, *а*, *и* характеризуются только одним формантом, звук *э* — двумя и звук *ы* — тремя. По данным [64, 66, 165], для хорошего распознавания гласных необходимо определить первые три форманта, а по [129] — два. Незвонкие согласные не имеют ярко выраженных формантных областей и различаются по моментам амплитуд нулевого (M_0), первого (M_1) и второго (M_2) порядка, характеризующих спектр в выбранной полосе [22]:

$$M_0 = \Sigma A_n, \qquad M_1 = \Sigma f_n A_n, \qquad M_2 = \Sigma f_n^2 A_n^2,$$

где A_n — амплитуда n-й полосы спектра, f_n — ее средняя частота.

Звук голосовых связок $h(t)$ имеет спектр частот четной и нечетной частей резонаторов, т. е.

$$H(j\omega) = \int_{-\infty}^{\infty} h(t)\, e^{-j\omega t}\, dt.$$

Резонаторы полости рта и носа в зависимости от образуемого звука речи имеют передаточные функции $G(j\omega)$, т. е. $G_а(j\omega)$, $G_о(j\omega)$, $G_у(j\omega)$ и т. д., соответствующие гласным *а*, *о*, *у* и т. д. Сигнал при выходе изо рта определяется уравнением

$$c(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(j\omega) H(j\omega)\, e^{j\omega t}\, d\omega.$$

Поскольку звук голосовых связок имеет спектр, почти одинаковый для всех людей, звуковой сигнал определяется только передаточной функцией резонаторов $C(j\omega) = G(j\omega)H(j\omega)$. Каждой фонеме соответствует своя эталонная передаточная функция. На основании этого явления разработан прибор узнавания человека по голосу [188], а также для получения информации о состоянии космонавта в ходе полета [1].

Каждый звук имеет свои оттенки. Тембр голоса любого человека различен в зависимости от свойств резонаторов и изменения основного тона во время разговора. Амплитуды и частоты формантов каждой фонемы могут изменяться в определенных пределах [87, 113, 131]. Это обстоятельство затрудняет конструирование прибора автоматического распознавания и синтеза, так как одни и те же концентрации энергии при одной частоте могут принадлежать разным фонемам [110].

О значении формантов имеются разные мнения. Одни исследователи считают, что их определение не может иметь решающего значения при разработке устройства распознавания речевых сигналов [79, 121], однако большинство авторов все же придерживается противоположного мнения [24, 67]. Так, если записать гласную *а* на магнитную ленту и подавить в ней некоторые области формантов, *а* превращается, например, в *у*.

Как известно, большинство фонем, в том числе и гласных, можно передавать шепотом, без участия голосовых связок. Возбуждающим воздействием является шум дыхания, который можно рассматривать как случайный стационарный процесс $n(t)$, имеющий спектральную плотность $N(\omega)$. Тогда спектральная плотность выходной величины, т. е. полученной фонемы, например *а*, будет равна

$$N_a(\omega) = |G_a(j\omega)|^2 N(\omega).$$

Как видно из этой формулы, основной тон информации о фонеме не несет, и поэтому при разработке аппаратуры автоматического распознавания звуков речи нет необходимости принимать его во внимание.

Передаточная функция фонем $G(j\omega)$ содержит как амплитудные, так и фазовые данные. Как известно, слух не реагирует на изменения фазовых сдвигов сложного сигнала, и поэтому при разработке аппаратуры автоматического распознавания этими сдвигами можно пренебречь; однако при синтезе сдвиг фаз между гармониками значительно ухудшает качество звучания речи [35, 41, 108, 147].

Для анализа речевых сигналов были разработаны динамические спектрографы (так наз. видеографы), дающие возможность получить трехмерное изображение речи: на оси абцисс — время, на оси ординат — частота и как третья координата — степень черноты точки на плоскости время—частота в зависимости от амплитуды сигнала при данной частоте [128, 137, 138] (рис. 2).

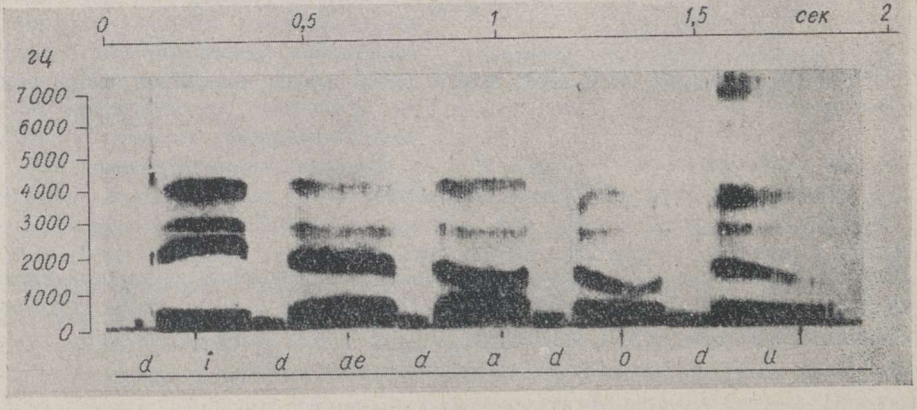

Рис. 2. Видеограф слогов. (Пример заимствован из работы [64]).

С точки зрения принципа работы видеографы разделяются на приборы с последовательным и параллельным анализом. Один из первых видеографов был разработан Х. Зундом [166]. Он дает также возможность фотографировать амплитуд-частотную зависимость на катодной трубке. Недостатком видеографа Х. Зунда является невозможность наблюдения и регистрации временных огибающих речевого сигнала.

Видоизменением видеографа является интервалограф [44]. В этом приборе получаются сигналы, амплитуды которых пропорциональны интервалам между переходами речевого сигнала через нуль.

По заказу Института языка и литературы АН ЭССР бывший Научно-исследовательский электротехнический институт СНХ ЭССР изготовил спектрограф с параллельным анализом. Комплект из 52 фильтров охватывает диапазон от 40 *гц* до 14 *кгц*. Спектрограмма анализируемого речевого сигнала получается одновременно на трех лучевых трубках, дающих возможность фотографировать и визуально наблюдать изменения спектра речевого сигнала.

Новый вариант спектрографа фирмы «Дженерал Электрик» (США) имеет 80 фильтров с шириной спектра 75 *гц* каждый, гетеродин, сканирующее устройство, узел логарифмирования амплитуд, камеру для фотографирования результатов анализа. Спектрограф может быть использован также для анализа песни птиц, шума в машинах, работы сердца и т. д. [186, 187].

При помощи видеографов установлено, что точность распознавания увеличивается не только путем определения расположения формантов и их интенсивностей, но и установлением скорости их изменения по частоте и уровням.

Речь представляет собой непрерывную функцию времени между паузами для дыхания и состоит из отдельных дискретных фонетических элементов — фонем. Фонемы суть разновидности звуков в зависимости от их произношения. Их всегда больше, чем звуков. В русском и английском языках их около 40, в эстонском — около 30, в немецком — 40 [47].

Проблема распознавания звуков речи может быть решена либо сравнением одной из фонем, слога или слова из соответственной совокупности, сохраненной в памяти машины, либо только по акустическим признакам. В первом случае в машине происходит сравнение признаков входного речевого сигнала с признаками сигналов, сохраненных в памяти машины, и входящий сигнал выдается как звук, имеющий наибольший коэффициент корреляции. Во втором случае сигнал выдается только в результате анализа.

Восприятие человеком речи можно разделить на три ступени. В первой — акустической — воспринимается ряд физических явлений и определяется комплекс параметров; во второй — фонетической — эти параметры сравниваются с эталонными параметрами в памяти и происходит первоначальное распознавание звука речи; наконец, в третьей — лингвистической — уточняется содержание полученной информации (например, к словам прибавляются окончания, которые диктор, может быть, вообще не произносил, и т. д.).

Автоматическое распознавание звуков речи базируется, главным образом, на решении первых двух ступеней. В общем случае машина для полного распознавания звуков речи должна содержать и устройства, хранящие информацию о языке.

Наиболее распространенными методами распознавания звуков речи являются анализы спектральных, временных и спектрально-временных характеристик.

Спектральный метод был впервые описан Л. Мясниковым [16, 17, 18], а позднее и другими [61, 129]. По его предложению звуковые колебания анализируются по парам фильтров, имеющим следующие полосы пропускания: 500—700 и 800—1000 *гц*; 1250—1500 и 4000—5000 *гц*; 650—750 и 5500—6500 *гц*; 1250—1500 и 400—500 *гц*. Выход из каждой пары детектируется и подается в противофазе на индикатор, стрелка которого либо остается в середине, либо отклоняется в одну или другую сторону. Фонемы разделяются только по частоте, независимо от величины амплитуд. Заменой индикатора на трехпозиционное поляризованное реле была получена точность распознавания до 75—80%.

К. Смиг в своем устройстве [160, 161] применял 32 фильтра и использовал принцип обострения формантных пиков путем определения разностей амплитуд в соседних фильтрах. Полученные сигналы подавались в сравнивающую систему, реагирующую только на сигналы с максимальной амплитудой. Определенная комбинация номеров каналов, т. е. формантных частот, соответствует определенной гласной фонеме. Но результаты оказались неудовлетворительными из-за того, что не исключались неопределенности, связанные с влиянием соседних фонем друг на друга и перемещением формантов в зависимости от изменения основного тона.

3. Вокодеры

На основании исследования формантов и спектрального метода анализа были построены полосные, формантные, сканирующие, гармонические и корреляционные вокодеры.

В полосном вокодере весь диапазон речевого сигнала анализируется в полосных фильтрах, имеющих либо по всему диапазону только равномерное, либо в нижних частотах (до 1000 *гц*) равномерное, а на высоких частотах (выше 1000 *гц*) логарифмическое разделение частот [104]. Выходы из каждого фильтра выпрямляются и являются параметрами анализируемого речевого сигнала.

В первом вокодере [59] речевой сигнал анализировался в десяти основных и двух вспомогательных фильтрах. Ширина полосы первого фильтра была 250, остальных — 300 *гц*, весь анализируемый диапазон составлял 2950 *гц*. Выход из каждого фильтра выпрямлялся, пропускался через вспомогательный фильтр низкой частоты (на $0 \div 250$ *гц*) и передавался через канал связи в синтезатор. Синтезатор представлял собой генератор шума с пределом частоты 250—3500 *гц*. Выход генератора соединялся с фильтрами генератора, имеющими такие же полосы частот, что и анализатор. Таким образом, выход фильтра генератора управлялся соответствующим выходом сигнала фильтра анализатора. Основной тон выделялся из сигнала до прохождения его через фильтры анализатора и пропускался для сглаживания через второй вспомогательный фильтр с полосой частоты 0—50 *гц*.

Поскольку в этом вокодере все высокие гармоники речевого сигнала не передаются, синтезированная речь звучит жестко и хрипло, но распознаваемость удовлетворительная.

В канале связи передается десять основных и два вспомогательных сигнала. Каждый канал требует ширину полосы около 25 *гц*, т. е. всего только 300 *гц*.

Если сравнивать некомпрессированные речевые сигналы с одинаковым объемом информации, но с разным отношением частотных и динамических диапазонов, то имеем

$$P_{c_2}/P_{n_2} = (P_{c_1}/P_{n_1})^{F_1 c_1 / F_2 c_2},$$

где P_{c_1}, P_{n_1}, P_{c_2} и P_{n_2} — мощности сигнала и шума на передающем и приемном концах канала связи; F_1 и F_2 — соответствующие полосы частот; c_1 и c_2 — соответствующие пропускные способности передачи количества сведений.

Полагая, что $c_1 = c_2$, т. е. пропускные способности обычной и вокодерной передачи одинаковы, то при $F_1 = 3000$ *гц* и $F_2 = 300$ *гц* имеем $P_{c_2}/P_{n_2} = (P_{c_1}/P_{n_1})^{10}$, т. е. теоретически вокодеры требуют в 10 раз большего динамического диапазона. В действительности такое требование преувеличено и, по данным ряда авторов [35, 63], если при передаче некомпрессированной речи достаточно иметь отношение сигнал/помеха около 30 *дб*, то для полосного вокодера с 10 каналами требуется отношение около 40 *дб* [161].

Вокодерные сигналы могут быть переданы либо непрерывно, либо в виде импульсов. Непрерывный сигнал имеет ширину спектра не более 15—50 *гц*. Динамический диапазон несколько меньше, чем у оригинальной речи, и не превышает 25—30 *дб*.

При импульсной передаче вокодерные сигналы квантуются по уровню и во времени. По частоте необходимо брать две пробы на 1 *гц*, а по амплитуде — через каждые 1—1,5 *дб*. При ширине спектра 25 *гц* и динамическом диапазоне ~ 16 *дб* канал должен обеспечить 200 *бит/сек* ($25 \times 2 = 50$ пробы; $2^4 = 16$ и $4 \times 50 = 200$) и суммарно при 10-канальном вокодере для общей пропускной способности канала требуется 2000 *бит/сек*.

Известно, что для передачи импульсов необходимо иметь частоту не менее 1—1,5 *гц* на 1 *бит/сек* и для передачи со скоростью 2000 *бит/сек* требуется, следовательно, канал с шириной частоты 3000 *гц*.

Согласно теории информации, речь может быть передана без искажений при выполнении условия

$$A \geqslant \frac{1}{3} D_{\text{ср}} F \qquad [\text{бит/сек}],$$

где A — пропускная способность канала; $D_{\text{ср}}$ — средняя величина эффективного динамического диапазона, F — ширина частотного диапазона речи. Если считать, что диапазон речевого сигнала составляет 5000 *гц*, средний динамический диапазон 30 *дб*, то $A \geqslant 5 \cdot 10^3$ [*бит/сек*]; следовательно, при приведенных данных нагрузка каналов связи уменьшается около 25 раз.

В полувокодере низкая частота речи (до 600 *гц*) передается без преобразования, а область высоких частот (выше 600 *гц*) анализируется и передается вокодером [69, 155]. По сравнению с вокодерами требуемый объем канала в этом случае увеличивался, но разборчивость однослоговых слов повышалась с 74 до 84%.

В сканирующем вокодере [175, 176, 178] речевой сигнал анализируется в 100 фильтрах, выход из которых детектируется и запоминается в конденсаторах. Уровень напряжения конденсаторов передается вращающимся переключателем со скоростью 30 раз в секунду в синтезатор, где звук восстанавливается при помощи управляемого мультивибратора. Разработаны также вокодеры, где основной тон не передается [34]; в этом случае получена разборчивость до 62,5%.

В импульсном вокодере [175] имеется 10 равномерно распределенных фильтров, выходы которых выпрямляются и управляют генератором импульсов так, что получается широтно-импульсная модуляция, причем количество импульсов для каждого форманта пропорционально по амплитуде и по частоте. По утверждению автора [175], этот вокодер обладает большой помехоустойчивостью.

Спектрально-временной метод распознавания отличается от спектрального тем, что выход фильтров сканируется еще по времени. В результате по каналу связи передаются фонемы в виде кодовых знаков [140, 142].

Дальнейшим развитием спектрального метода является формантный метод. Он состоит из определения наличия данного форманта в данной полосе фильтров [67—69, 75] (рис. 3). В вокодерах этого типа выделяется до четырех формантов. Усовершенствованные формантные вокодеры, в которых анализируются интенсивности, а также корреляции между частотами и амплитудами формантов, были разработаны несколькими авторами [3, 50, 64]. Формантные вокодеры, в которых учитывались моменты первого и второго порядка, дали лучшие результаты, особенно для определения и синтеза согласных, по сравнению с вокодерами, названными выше [43, 63, 85]. Предлагалось определять первый формант и в диапазоне 250—850 *гц* и в диапазоне 300—1200 *гц*. По [85], достаточно иметь верхний предел фильтра определения первого форманта не выше 1000 *гц*. Второй формант определяется в пределах 900—2300 *гц*. Поскольку в таком вокодере частота формантов перекроется, то предложено сначала определять расположение первого форманта. Если он находится в пределах 700—900 *гц*, то частота второго форманта не ниже 1400—1800 *гц*; если же первый формант находится значительно ниже 700 *гц*, то частота второго форманта лежит в пределах 700—900 *гц* и соответственно необходимо перестроить фильтры. Если передается еще третий формант, то он определяется в пределах 2100—3500 *гц*, а четвертый — в области 4000—

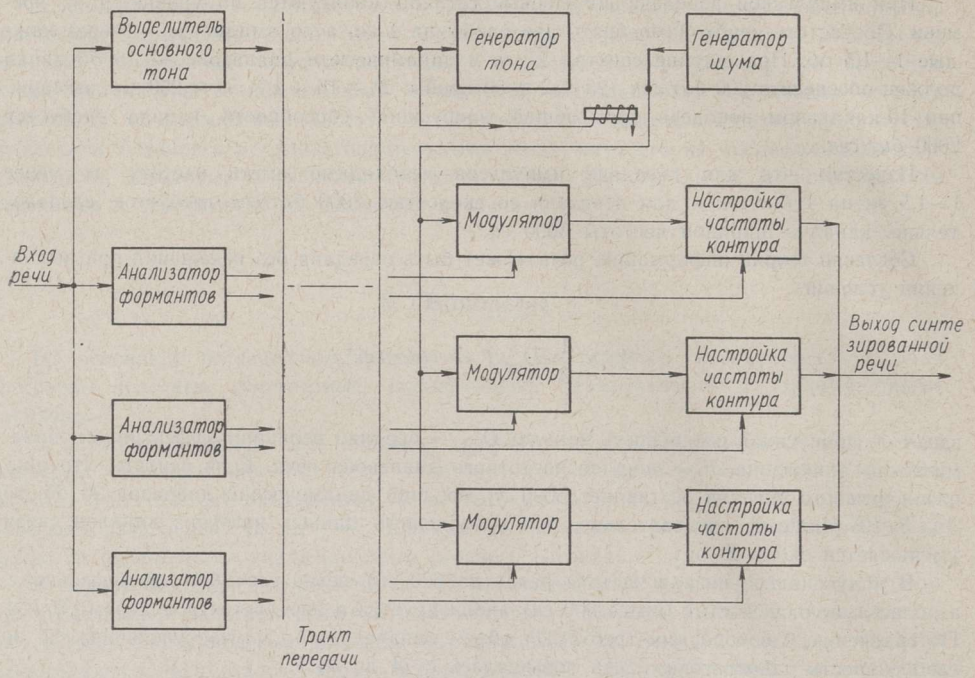

Рис. 3. Блок-схема формантного вокодера.

6000 *гц*. Моменты определяются дифференцированием спектра. Если основной тон передать двумя параметрами (частоты и уровня) и четыре форманта тремя параметрами, то необходимо передать всего 14 сигналов. Если же передать моменты, а также динамические показатели формантов как скорость и знак изменения частот и уровней, то число передаваемых сигналов увеличивается.

В сканирующих вокодерах имеются также анализирующие фильтры, но сигналы передаются к экспандеру последовательно во времени [177]. Если в формантных вокодерах речевой сигнал анализируется в отличие от полосовых вокодеров только в частотах, соответствующих расположению формантов в речевом сигнале, то гармонические вокодеры анализируют речевой сигнал полностью, определяют коэффициенты Фурье и члены ряда (кроме постоянной составляющей) передаются по каналам связи двумя параметрами. На приемном конце эти параметры управляют либо генератором дискретного спектра, либо генератором шума, а иногда и обоими [20, 21, 27]. По принципу работы гармонические вокодеры мало отличаются от полосовых. В обоих определяются ординаты спектра, но только в полосовых они передаются без преобразования, в гармонических же передаются их линейные комбинации. Разница заключается в синтезе на приемном конце: в полосовых вокодерах параметры задаются фильтрами, имеющими такие же полосы, как анализаторы, гармонические же вокодеры имеют на приемном конце фильтры, соответствующие разложению в ряд Фурье.

Корреляционный метод анализа основан на связи между автокорреляционной функцией $R(\tau)$ и энергетическим спектром сигнала $S(\omega)$ [23, 33, 154]:

$$S(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-j\omega\tau} d\tau = 2 \int_{0}^{\infty} R(\tau) \cos \omega \tau d\tau.$$

Корреляционный метод позволяет избежать влияния эффектов фазовых сдвигов в синтезированной речи, которые появляются в полосных, формантных и гармонических

вокодерах из-за наличия комплексного сопротивления полосовых фильтров синтезатора. Таким путем достигается 10-кратная компрессия [153, 154].

Для улучшения распознаваемости фонем спектрально-временным методом предложено все фонемы предварительно разделить по их характерным признакам [94, 95]. Электронная бинарная система [43, 45, 185] разделяет звонкие и глухие при помощи фильтров — звонкие имеют основной тон, глухие нет. Следующим шагом звонкие шумовые отделяются от нешумовых наличием или отсутствием на выходе фильтров первого форманта. Глухие разделяются на взрывные и шипящие при помощи разниц их амплитуд. Блок разделения звонких звуков на шумовые и нешумовые работает с точностью до 95%, блок разделения гласных на высокие и низкие — до 98%, а низкие на диффузные и компактные — до 94%.

Сравнительные данные об объемах каналов связи приведены в табл. 1 [159].

Таблица 1

Метод кодирования	Необходимый объем канала, *бит/сек*
Дискретная форма речевого сигнала	30000 *
Фонемный	60
Словесный (120 слов в минуту):	
а) словарь из двух слов	2
б) словарь из 8000 слов	26
Вокодер	2000
Телетайп (120 слов в минуту)	75

* Из расчета, что диапазон речевого сигнала составляет 3000 *гц*.

Улучшению технических показателей вокодеров посвящен ряд работ [49, 118, 164]. Так, разработаны цифровые вокодеры, где вместо возбуждения синтезатора голосом, как это было сделано в первых вокодерах, возбуждение происходит основным тоном, который в свою очередь включает специальный генератор. В синтезаторе используется редуцирующее устройство, которое устраняет помехи, возникающие вследствие варьирования основного тона [171]. Разработана новая система компрессии, не требующая выделения основного тона [76]. Использованием многоканального модулятора спектр передаваемого сигнала снижен до 1,4 *кгц* [77]. Предложена схема вокодера, позволяющая значительно уменьшить объем канала связи (до 94 *бит/сек*); работа вокодера основана на использовании памятного устройства, сохраняющего все фонемы, причем каждая фонема имеет свой код и «вызывается» при помощи цифровых сигналов [47].

Другое предложение для уменьшения объема канала связи основано на использовании слоговых синтезаторов. Из анализатора получается код слогов в виде цифр, которые передаются в синтезатор, и выбор необходимого слога производится по этим кодовым цифрам. По этому принципу изготовлен синтезатор с объемом 200 слогов [126].

Электронно-аналоговый синтезатор EVA воспроизводит звуки речи по кривым, нарисованным токопроводящими чернилами на ленту или барабан [96]. Передача такой речи может быть осуществлена при частоте в 30 раз меньшей, чем при обыкновенном телефонном переговоре. Разборчивость слов достигается до 75%. По утверждению автора [96], разборчивость может быть доведена до 85%, а уплотнение по частоте увеличено до 1000. Такая передача требует очень незначительной мощности, что крайне важно при установлении связи с космонавтом.

4. Специальные приборы распознавания речевых сигналов

Спектрально-временной метод был в подробностях изучен Х. Олсоном и Х. Беларом [122, 123] и Ж. Дрейфус-Графом [57, 58], которые разработали пишущие машинки,

управляемые устными командами. Первая машинка Х. Олсона и Х. Белара печатала слова из репертуара десяти однослоговых слов, имеющихся в памяти машинки; третья имела уже словарь из 100 однослоговых слов (рис. 4). По мнению авторов, машинка

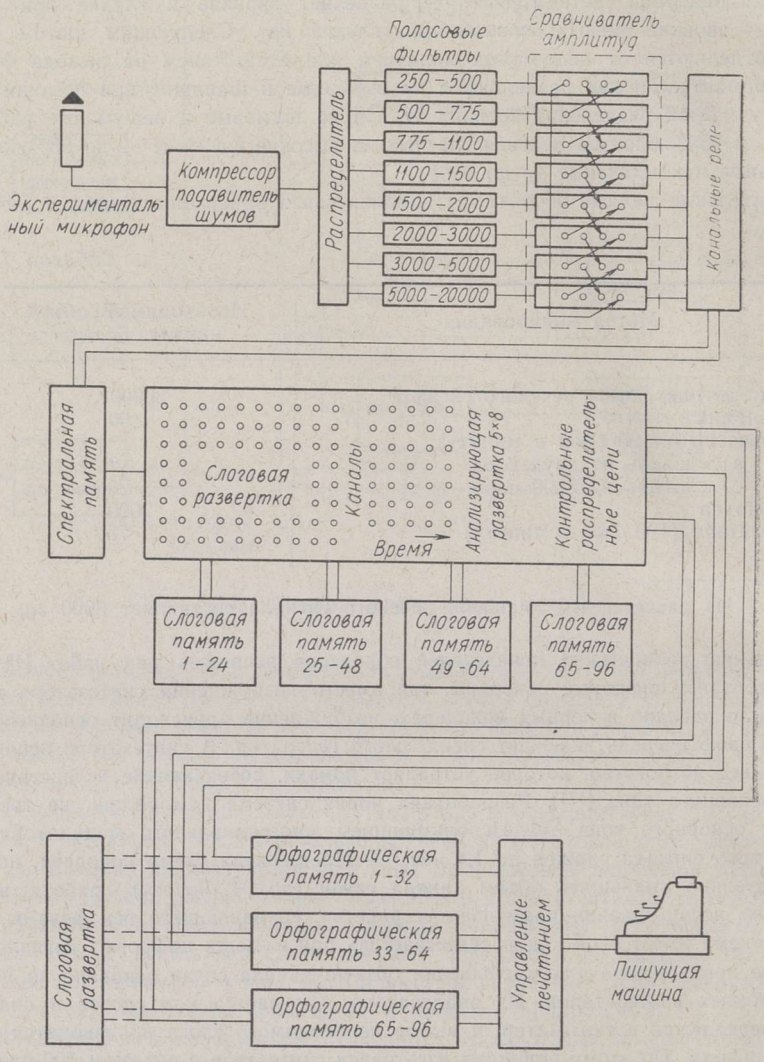

Рис. 4. Блок-схема фонетической пишущей машинки Х. Олсона—Х. Белара.

с памятью в 1000 звукосочетаний достаточна для практического использования [124] (очевидно, для русского и эстонского языков этого требования также достаточно, так как в этих языках фонетическое произношение и печатная форма мало отличаются друг от друга). Последняя модель их пишущей машинки состоит из 8 полосовых фильтров, 8 амплитудно-сравнивающих детекторов, слоговой и орфографической памятей, контрольных устройств и пишущей машинки. Фильтры охватывают диапазон от 250 до 20000 *гц*. Выход из каждого канала подается в устройство, где происходит обострение максимумов сравнением уровней выхода соседних каналов и определением второй производной кривой речевого сигнала. Квантование входного сигнала по времени происходит через 0,2 *сек*, а после его выхода из фильтров в память через

0,04 *сек* (всего также 0,2 *сек*), а по амплитуде по трем уровням. В схеме сравнения происходит сравнение выходной кривой речевого сигнала до и после квантования, т. е. через 0,04 *сек*. Если изменения амплитуды сигнала в это время не происходит, то сигнал не достигает памяти. В памяти имеется 256 разных знаков, в том числе и пауза. Для передачи результатов анализа требуется 8-значный код, что при скорости произношения 10 фонем в секунду требует скорости 80 *бит/сек*. Точность работы машинки 92—94%, если она настроена на определенного диктора, и значительно хуже при случайном дикторе.

Х. Олсон и его сотрудники разработали устройство, в котором происходит распознавание, перевод, печатание и синтез речи [125]. Память машины состоит из 4 английских, 8 французских, 4 немецких и 4 испанских слов. Скорость работы машины 60 слогов в минуту. Точность работы при определенном дикторе 96—98%.

В четвертой, находящейся еще в стадии построения модели фонетической пишущей машины Ж. Дрейфус-Графа [58], по заявлению автора, исключены особенности речи разных дикторов. Как в предыдущих, так и в этой машинке названного автора память отсутствует, и поэтому точность ее работы в значительной мере зависит от четкости произношения дикторов. Анализ спектра речи происходит в 10 фильтрах (400—3200 *гц*), кроме того, имеются еще фильтры диапазоном 150—300 и 4600—6000 *гц*. В каждом спектральном канале имеются детектор и фильтр низкой частоты для определения субформантов (0—30 *гц*) и квантизатор. Определяется еще скорость изменения огибающей речевого сигнала. В итоге речевой сигнал разлагается на 48 сигналов, каждый из которых квантуется еще по времени через $^1/_{15}$ *сек* и по трем уровням. Суммарно имеется информация в объеме 1440 *бит/сек*, алфавит машинки имеет 30 букв. Данные о точности работы машинки еще не опубликованы.

Пишущая машинка М. Кальфаяна [100] имеет устройство приведения речевого сигнала к одной частоте основного тона при помощи генератора, управляемого основным тоном входного сигнала. Имеются фильтры, выходы которых после выпрямления сравниваются при помощи электронного реле с данными, сохраненными в памяти; машина печатает фонемы в фонетическом, а не в печатном обозначении. Никакой коррекции ошибок произношения не производится. Подробности работы машинки еще не опубликованы.

Механический распознаватель речевых сигналов, разработанный в Лондонском университете, имеет в отличие от предыдущих лингвистическую память. Как указывает автор [51], результаты анализа могут быть использованы для управления пишущей машинкой, но он считает, что успех будет не особенно велик. Распознаватель может различить 4 гласные и 9 согласных. Различение происходит по максимумам напряжений двух из 18 фильтров, охватывающих диапазон речевого сигнала от 160 до 8000 *гц*. Установлено, что, например, гласному *i* соответствует максимум произведений выходов фильтров 250 и 3200 *гц*, согласному *m* — 200 и 320 *гц*. Согласные, имеющие спектры почти одинакового состава, различаются далее либо по длительности, либо по уровню напряжения. Лингвистическая память содержит сведения о вероятностях следования двух фонем друг за другом. Устройство состоит из нескольких потенциометров, и информация вводится в матрицу потенциометров положением их скользящего контакта. В акустическом устройстве определяются форма спектра, длительность и интенсивность звука и наличие основного тона. В лингвистическом устройстве заложены типовые комбинации имеющихся в речи фонем, а комбинации, которые в речи не существуют, запрещены. Таким образом, распознавание происходит в акустическом устройстве, а поправка, т. е. улучшение узнаваемости, в лингвистическом. Эксперименты с 200 словами дали точность распознавания 72% (для любого голоса 45%).

Корреляционный метод распознавания основан также на умножении выходов фильтров. Разработанные в лаборатории Белла (США) устройства предназначены для распознавания 10 цифр устного выбора желаемого абонента [50]. Воспринятый речевой спектр умножается на типический спектр фонем, после чего находится их среднее значение. Максимальное значение умножения соответствует той фонеме, которая более похожа на принятую. Устройство состоит из комплекта, включающего 10 реле, причем

выходной импульс дает всегда только одно реле. Точность распознавания составляет 97—99% при определенном и 50—60% при случайном дикторе.

В Японии разработано устройство для распознавания изолированно произнесенных 10 цифр. Этот процесс производится по восьми признакам (число звонких интервалов в слове, положение первого и второго формантов в начале первого звонкого интервала, их положения через 100 *мсек* после начала первого звонкого интервала и т. п.), имеющим от 2 до 13 уровней. Напряжения уровней этих восьми признаков поступают на матричную схему, в которой производится вычисление условных вероятностей всех цифр. Наибольшая условная вероятность соответствует произносимой цифре. При произношении 1000 слов одним диктором получено 99,7% правильных результатов, а при 20 дикторах (мужчинах) — 97,9% [189].

Результаты экспериментального исследования по определению разных признаков произносимых цифр на русском языке приводятся в [30].

Временной метод распознавания основан на использовании клиппированной речи. Этот метод был изложен в работах И. Ликлайдера [112] и вслед за ним в работах других ученых [145, 150, 174], в том числе и советских [12]. Г. Цемель [28] использует клиппированную речь для различения некоторых согласных. По его данным, точность распознавания для звуков *п* и *т* составляла 95, а для звука *к* — 75%.

А. Райс [145] анализировал согласные, основываясь на их хорошей отличимости на слух. Эксперименты были проведены для выяснения возможности ввода данных в устном виде в машину автоматического перевода. Было установлено, что зависимость числа импульсов клиппированной речи от времени для различных гласных, произнесенных разными дикторами, почти линейна. Результаты опытов с тремя дикторами приведены в табл. 2. Предварительное дифференцирование сигнала отсутствовало.

Таблица 2

Диктор	*а*	*о*	*и*	*і*
	имп/сек			
1-й	558	450	350	283
2-й	533	417	317	—
3-й	491	384	292	218

И. Тоффлер предложил метод выделения из речевых сигналов основного тона путем использования нелинейных элементов и метода клиппирования [172].

Клиппированная речь исследовалась также в университете Киото [150]. Анализу были подвергнуты как гласные, так и согласные. Согласно примененной здесь методике речевой сигнал после усиления до выбранного уровня подается на вход каскада одновибраторов, выходом которых являются импульсы, соответствующие моментам изменения направления прямоугольного клиппированного речевого сигнала. Каскад одновибраторов, длительность импульсов которых в отдельности имеет разные значения, классифицирует клиппированный речевой сигнал на 14 значений от 0,59—1,11 до 22,55—32,22 · 10^{-4} *сек*. В каждом канале число импульсов определяется счетчиком. Анализ гласных производится по распределению временных интервалов $W(t)$ и одномерному распределению вероятности $W_1(t)$ по формулам

$$W(\tau_{m_i}) = \frac{1}{\Delta\tau_i} \frac{\nu_i}{\sum\limits_{i=1}^{14} \nu_i} \quad \text{и} \quad W_1(\tau_{m_i}) = \frac{1}{\Delta\tau_i} \frac{\tau_{m_i}\nu_i}{\sum\limits_{i=1}^{14} \tau_{m_i}\nu_i} \qquad (i = 1 \ldots 14),$$

где τ_{mi} — средний временной интервал i-го канала: ν_i — число интервалов в i-ном канале; $\Delta\tau_i$ — разность по времени между верхней и нижней границами i-го канала.

Пишущая машинка, основанная на анализе клиппированной речи, построенная в том же университете [149], имеет память в объеме 200 слогов. Подробности работы машинки еще не опубликованы.

Устройство автоматического различения из 20 произносимых слов (цифры от 0 до 9, плюс, минус, пробел, вперед, назад и др.), базирующееся на клиппировании предварительно дифференцированного речевого сигнала, разработано в Академии наук Грузинской ССР [13]. Утверждается, что устройство способно различить до 20 речевых команд при настройке его на определенный голос и порядка пяти от многих голосов.

Портативный электронный арифмометр SHOEBOX, пока единственная серийная модель, реагирует на произнесенные вслух 10 цифр (от 0 до 9) и 6 команд: плюс, минус, сумма, частная сумма, ошибка, сброс. Слово «ошибка» останавливает устройство и гасит все произведенные операции. Распознавание происходит по месту расположения первой фонемы, ударной гласной, последней фонемы и временной огибающей положительных и отрицательных пиков кривой слова. При разных дикторах работает нестабильно [115].

Из описанных трех пишущих машинок [58, 123, 149] самой перспективной, согласно материалам Стокгольмского семинара 1962 г. [139], считается машинка Ж. Дрейфус-Графа, не имеющая ограничительного узла — памяти [32, 141].

Приведем еще некоторые данные о специальных приборах анализа, распознавания и использования речевых сигналов.

Для анализа и распознавания сигналов звуковой частоты разработано устройство, действие которого основано на резонансной механической системе, состоящей из стекловолокон различной длины и диаметром до 0,05 *мм*. Одним концом эти волокна закреплены на специально профилированном основании, другой может свободно колебаться под действием звуковых волн. При помощи световых лучей колебания этих свободных концов проектируются через эталон на фотодиоды. Элемент состоит из 2000 волокон общим объемом 1,6 $см^3$, анализирует диапазон частот от 30 до 20000 *гц* с разрешающей способностью около 10 *гц*. В эталоне сохраняются сигналы прототипов. Фотоэлемент интегрирует все количество света, и чем больше данный сигнал соответствует сигналу в эталоне, тем больше этот ток. Если сила тока превышает заданный порог, звуковой сигнал распознается. Поскольку эталон также может быть изменен в зависимости от полученной информации, то описываемый элемент владеет признаками самообучения. Устройство распознает только короткие слова, произносимые конкретным голосом. Этот прибор был построен для попытки установить связь с дельфинами. Установлено, что «речь» дельфинов образуется короткими сигналами (~0,1 *сек*) при частоте от 5 до 10 *кгц* [180, 181].

Для гармонического анализа периодических функций, заданных в виде графиков или таблиц, разработан анализатор гармоник электромеханического типа, позволяющий одновременно получить пять пар коэффициентов ряда Фурье с точностью 0,3% от максимального значения анализируемой функции [4].

Построен прибор, в котором для распознавания команд (цифр) используются визуальные данные о движении губ. С обеих сторон губ установлены осветители с направляющими рефлекторами. Перед губами расположено снабженное собирающим рефлектором фотосопротивление, которое включено в одно плечо сбалансированного моста. Снимаемое с моста напряжение усиливается, подается в дифференциальный усилитель и оттуда в самопишущий прибор. Осветители и фотосопротивление с рефлектором прикреплены к голове говорящего. Распознаваемость десяти цифр для конкретного диктора составила 91%, для двух разных дикторов — 78,3%. Определением еще одного параметра — скорости движения потока воздуха около губ — удалось повысить распознаваемость при одном дикторе до 100, для двух — до 81% [84].

Теоретически разработана система соленоидов, при помощи которой можно распознавать разные коды на 24000 английских слов длиною до 16 букв, однако для распознавания слов она еще не построена. По сути дела эта система представляет собой устройство памяти, в которой сохраняются кривые речевых сигналов в цифровом виде

и можно получить моментальные значения корреляционной функции речевого сигнала [38, 134].

На основе использования релейных систем создано устройство для преобразования цифровой информации в речевую. На дорожках магнитного барабана записаны звучания названий цифр от 0 до 9 и слова «вольт», «секунда», «степень». Адрес соответствующей дорожки барабана выдается кольцевым счетчиком, а выбранная запись поступает в звуковой усилитель [144].

Нейронная сеть и слуховой аппарат человека моделированы несколькими авторами [19, 74, 83, 90]; разработана также корреляционная теория слуха [101]. На базе модели разработано устройство для распознавания звуков и отдельных слов. Устройство состоит из выходного усилителя, системы фильтров и блока логических и решающих схем, моделирующих наружное и среднее ухо, внутреннее ухо и нервные сети соответственно. Для разделения звонких согласных *b*, *d*, *g* и глухих *p*, *t*, *k* использовались дифференцирующие логические схемы, однако полного электронного аналога еще не имеется [116].

Проведены эксперименты «понимания» речи, не слушая ее. Информация о фонеме передавалась через руку с помощью 24 вибраторов, соединенных с выходами функциональной модели органа слуха. Цифры от 0 до 9 были записаны на магнитную ленту. В случае неправильного ответа об этом сообщается обучающемуся, и подача данной цифры повторяется. Через 2—3 часа обучения получено 85% правильных ответов. Этот эксперимент имеет большое значение для глухих [190].

Человеческий голос не имеет симметрии относительно оси перехода, как, например, спектр шума. Эта особенность — «асимметрия огибающей» — была использована для создания соответствующего предохранителя, выключающего мощную машину, например станок, по крику рабочего [163].

Разработан анализатор речевых сигналов, состоящий из 54 фильтров гауссового типа [80]. До 1000 *гц* фильтры имеют ширину 70 *гц*, и при частоте выше 1000 *гц* ширина их увеличивается по сравнению с предыдущими на 6,5%. Выход каждого фильтра выпрямляется и квантуется; квантованные токи можно проследить визуально и после соответствующей обработки подать в ЭВМ [169].

Разработан анализатор речевых сигналов, состоящий из 96 фильтров и охватывающий диапазон частот от 30 до 8000 *гц*. Сигналы нормализуются, детектируются, определяются до 4-й производной, а выход анализатора соединяется с трехлучевым осциллоскопом. Разрабатывается автоматизация процесса их распознавания [53].

Решение проблемы выделения основного тона имеет большое значение как для лингвистов, так и для изготовления вокодеров и для решения вопроса распознания речевых сигналов вообще [78]. Разработано несколько вариантов специальной аппаратуры [42, 56, 147] и устройств, отличающихся узлами определения автокорреляционной функции сигнала для выделения максимумов пиков [72], шириной формантов [62, 170] и другими дополнительными узлами [143, 184]. Так, в [184] предлагается изменить фазу выходов из N фильтров на 90°. Эти выходы рассматриваются как многомерные векторы, изменяющиеся по времени. В случае наличия периодичности в сигнале этот вектор образует замкнутую кривую. Используется 5 фильтров шириной 120 *гц*, охватывающих диапазон частот от 300 до 900 *гц*.

Для выделения основного тона в речевом сигнале разработано устройство, состоящее из элементов с нелинейной характеристикой и незначительными постоянными времени [151]. Другая система анализа частот формантов и основного тона работает по принципу следящего фильтра с предварительным переносом спектра анализируемых частот [82] или генерированием сигнала и ручной подгонкой его до совпадения с анализируемым сигналом [179].

5. Универсальные вычислительные машины как средства исследования и распознавания речевых сигналов

Использование универсальных вычислительных машин для исследования речевых сигналов началось уже после их создания [48, 70, 88] и расширяется ежегодно. Их применяют в связи со спектральным анализом речевых сигналов [5, 37, 46, 99], с распознаванием и синтезом выбранной комбинации фонем [38, 88, 91, 116], цифр [52, 71, 152], с выделением основного тона [46, 72, 81, 120] и определением параметров форматов [127, 167, 182], с исследованием работы вокодеров [132], с нахождением возможностей ввода речевого сигнала в вычислительную машину с целью математического моделирования систем связи [15, 46] и т. д. В [120] приводится метод разложения речевого сигнала в ряд Фурье и определения его константы. Амплитуды каждой последовательности частоты логарифмируются и их анализируют во втором спектральном анализаторе. Выход этого анализатора представляет собой логарифм спектра мощности и имеет пиковые значения в случае анализа звонких фонем, но не имеет их в случае анализа глухих и других фонем, у которых основной тон отсутствует. Поскольку временные изменения частот речевых сигналов вызывают периодические пульсации спектра амплитуд, преобразование Фурье спектра дает частоту пульсации, обратно пропорциональную частоте основного тона речи.

Проведен спектральный анализ группы гласных звуков, предварительно сегментированных, с применением специального устройства, которое производило цифровое кодирование анализированных речевых сегментов. Эти данные вводились в ЭВМ и проводился шаговый синхронный анализ по методу Фурье [37].

В [135] предлагается новая техника для измерения частоты и ширины формантов речевого сигнала, основанная на теории Фанта [24]. Задается вид уравнений спектра

$$f(t) = a_0 + \sum_{i=1}^{N} e^{-\pi B_i t} (a_i \cos 2\pi F_i t + b_i \sin 2\pi F_i t),$$

где $N = 3$ и 4 (число формантов).

Дается методика для нахождения числовых значений коэффициентов a_0, a_i, b_i, $F_i (1 \leqslant i \leqslant N)$ при помощи ЭВМ по методу наименьших квадратов (B_i — ширина i-го форманта и F_i — его частота). На основании экспериментальных материалов анализа трех слов (bought, bottle и beet), произнесенных двумя лицами по два раза, утверждается, что два первых форманта для ə, *i* и *a* найдены вполне надежно.

Для расчета энергетического спектра речевых сигналов существуют, кроме разложения его на ряд Фурье, и другие методы, например разложение спектра на полиномы или представление его в виде марковского процесса [93]. В [5] приводятся результаты корреляционно-спектрального анализа, в [10] — оценки погрешностей метода спектрального анализа, а в [7] — оценки частоты квантования спектра речи при корреляционном и спектральном анализе его на ЭВМ.

В [148] утверждается, что проблема распознавания различных образцов, включая и речевые, сводится к задаче нахождения класса, к которому принадлежит данный сигнал, если известно общее число классов сигналов, куда этот сигнал входит. Эта задача решается путем минимизации некоторой функции риска, в результате чего находятся оптимальные правила, используемые для решения задачи распознавания выходных сигналов электронного анализатора, представляющего собой модель ушной улитки.

Предложен метод представления модели форманта в виде n-мерного вектора, каждый компонент которого представляет собой одну дискретную величину форманта в данный момент времени [40]. Таким образом, размеры n-мерного пространства определяются умножением дискретных величин параметров формантов на моменты наблюдения. ЭВМ разрабатывает эти данные двумя шагами, названными автором процес-

сами обучения и распознавания. Результаты анализа каждой из 10 фонем (первые буквы алфавита), произнесенных 10 дикторами, представлены в виде матрицы [40].

Для распознавания речевых сигналов предлагается разработать программу для вычисления на ЭВМ энергии и огибающей речевого сигнала за выбранный промежуток времени, частоты перехода сигнала через нуль и распределения промежутков между нулями клиппированной речи в течение этого промежутка времени, автокорреляционной и взаимной корреляционной функции а также спектрального анализа речевого сигнала [9]. С этой целью для ввода звуковой информации в ЭВМ разработан преобразователь аналог-цифра при 8-разрядном отсчете двоичных чисел [6].

В [53] речевой сигнал анализируется в 30-канальном полосном анализаторе. С помощью ЭВМ определяются частоты F_1, F_2 и F_3 через 10 *мсек*. Участки речи, в течение которых частоты формантов не меняются, во внимание не принимаются. Эксперимент по опознанию 10 моносиллабических слов тремя дикторами по 3 раза дал 100% правильных ответов. Также успешным было опознание дикторов по их голосам.

Проблеме выделения основного тона посвящено также много работ. Кроме специальных аппаратов, указанных выше, такие работы ведутся и на ЭВМ в большинстве случаев совместно с определением других параметров речи. В [81, 182] разработана программа, с помощью которой предварительно обработанный в анализаторе корреляционного типа речевой сигнал проходит преобразователь аналог-цифра, вводится в ЭВМ для определения параметров основного тона и других показателей речевого сигнала, как-то: частота и амплитуда первых трех формантов, мгновенная мощность сигнала и скорость изменения всех этих величин. Результаты выделения основного тона сравниваются с данными, полученными в [109], в которой при исследовании речевых сигналов на ЭВМ определение основного тона и скорости его изменения при произношении отдельных слов и слогов производится в общем случае автоматически, а более точное определение ручным измерением расстояний между пиками осциллограмм речевых сигналов. Отмечается хорошее совпадение результатов измерения.

В [167–169] рассмотрен процесс выделения частот формантов и способ их определения в односложовых словах японского языка на ЭВМ. Вначале речь предварительно анализируется в системах фильтров, которые охватывают диапазон частот от 200 до 5900 *гц*, затем кодируется 8-разрядным двоичным словом и посылается в магнитную ленту ЭВМ. В ЭВМ из речи выделяются формантные частоты, определяются скорость изменения формантных частот и моменты второго порядка вблизи средней частоты, производится сегментация фонем и фонемная классификация спектра. Утверждается, что предполагаемая методика позволяет безошибочно выделить изолированно произнесенные фонемы, почти безошибочно выделять и распознавать звуки и глухие согласные.

Для автоматизации обмена информацией между человеком и складом создана система, в которой звуковая речь анализируется по энергетическим признакам, преобразуется в двоичные электрические сигналы с помощью устройства, содержащего частотные фильтры, преобразователи аналог-цифра и дешифраторы, кодируется на перфокартах и передается в ЭВМ. Абонент обращается к ЭВМ по телефону и получает ответ на разных языках (французском, английском и др.) [107]. Аналогичная система (на ЭВМ IBM-7770) с объемом памяти в 60 слов может давать ответы на 750 поступающих одновременно телефонных запросов относительно цен на бирже [8].

Своеобразное направление исследования речевых сигналов на ЭВМ, вызванное поисками путей сокращения количества информации о звуке речи, проявляется в применении метода «анализ-синтез», предложенного еще К. Стивенсоном и др. [86, 98, 127]. В [127] первоначальное определение параметров речевых сигналов производится путем сравнения входного спектра со спектром, образующимся в системе в результате комбинирования 6 кривых для первых формантов и 6 для вторых формантов, т. е. всего 36 эталонных спектральных кривых. При образовании эталонных спектров изменяются 8 параметров: частота и ширина полос первых трех формантов, частота четвертого форманта и положение нуля в спектре источника. Утверждается, что разработанный метод позволяет достаточно хорошо исследовать речь и получать исходные данные для разработки аппаратуры распознавания речи.

В другом предложении речевой сигнал непосредственно, без дополнительных аппаратов, подвергается математическому анализу в ЭВМ и аппроксимируется 30 ортогональными функциями в виде экспоненциально затухающего ряда Фурье. При изменении звука меняются числовые значения коэффициентов, сами же функции остаются неизменными. Полученные символы используются для синтеза [55].

Аналогичные работы ведутся также в Японии [91, 98] и Польской Народной Республике [97].

Уменьшения влияния субъективных свойств дикторов на результат машинного распознавания можно в некоторой мере достигнуть путем применения принципа самонастройки и самообучения. Если же разговаривает много людей, как в обыкновенной беседе, то этот принцип, конечно, малоэффективен. В [173] описываются результаты работы по распознаванию образцов с применением этих методов совместно с методом «анализа-синтеза». Из 20 признаков, например, можно получить чрезвычайно большое количество комбинаций, но многие из этих признаков несущественны и поэтому создаются субматрицы, в которые входят только существенные признаки. Степень важности каждого признака и их комбинации друг с другом определяются в ходе самообучения, т. е. синтезированием подходящего «словаря». Вначале машина генерирует все признаки и, анализируя связи между отдельными ячейками, определяет вес каждой из них, потом генерирует новые связи и определяет их веса. Методом сравнения происходит распознавание. Количество правильных ответов при распознавании известных стилизированных портретов и спектрограмм речи было 80—100%, а для неизвестных 60—100%.

Алгоритм машинного распознавания с применением элементов самонастройки разработан также в [183]. При памяти объемом в 18 слов точность различения такой самонастраивающейся системы переработки информации, подаваемой в машинку на четырех языках, составляла 96%, а в случае 20 слов — 86%; система удовлетворительно различала голоса, относительно которых не имела предыдущего опыта.

Разработана программа, при которой в объеме ограничительной памяти (числа от 0 до 9, плюс, минус, равняется, скобки и т. д., всего 83 слова) вычислительную машину «учили» распознавать эти слова как для определенного, так и для случайного диктора (с ухудшенным результатом), выполнять приказы арифметического действия с произнесенными цифрами, выдавать результат в печатном виде и переводить на другой язык [132].

Вычислительная машина для распознавания речевых сигналов обладает не только тем преимуществом, что она дает возможность использовать частичную информацию и алгоритмы, разработанные для этой машины, для перевода с одного языка на другой, но и тем, что большой объем памяти и высокая скорость действия позволяют разлагать речевой сигнал для анализа в очень коротких участках как по амплитуде, так и по частоте, и проводить быстрое сравнение с эталонными данными. Но, с другой стороны, для достижения универсальности распознавания, т. е. независимости результатов анализа от субъективных свойств дикторов, она не должна содержать элементов сравнения; кроме того, большие вычислительные машины дороги и не могут быть мобильными. Поэтому несмотря на некоторые преимущества по сравнению со специальными машинами, последние, т. е. машины непосредственного анализа, кажутся более перспективными при окончательном решении машинного распознавания речевых сигналов и использовании их в различных системах управления и связи.

В настоящей статье затронута только наиболее общая часть работы в области машинного распознавания и совсем не затронут вопрос о синтезе речи. Главным фактором понижения точности работы аппаратов распознавания речи является непостоянство показателей формантов, во многом зависящее от индивидуальностей говорящих, и трудность сегментирования. Но результаты исследования уже нашли применение в технике связи, а также в военном деле. Так, в США разработаны вокодеры, применяемые в авиации [2, 133]. По мере совершенствования аппаратов область их применения, несомненно, увеличится, в результате чего в кибернетических системах появятся новые элементы, которые будут реагировать на устные команды человека без промежуточного кодирования.

ЛИТЕРАТУРА

1. Анализ поведения человека по его речи, Электроника, **37**, № 19 (1964).
2. Армия закажет миниатюрный вокодер, Электроника, **37**, № 15 (1964).
3. Варшавский Л. А., Литвак И. М., Исследование формантного состава и некоторых других физических характеристик звуков русской речи, Пробл. физиол. акустики, **3** (1955).
4. Василенко А. Т., Денисов Ю. Н., Электромеханический анализатор гармоник, Приборы и техника эксперимента, № 6 (1963).
5. Волошин Г. Я., Спектральный анализ речевых сигналов с помощью ЭВМ, Сб. тр. Ин-та математики СО АН СССР, «Вычислительные системы», вып. 10, Новосибирск, 1964.
6. Волошин Г. Я., Преобразователь аналог-цифра для ввода речевых сигналов в ЭВМ, Сб. тр. Ин-та математики СО АН СССР, «Вычислительные системы», вып. 10, Новосибирск, 1964.
7. Волошин Г. Я., О частоте отсчетов случайной функции при корреляционно-спектральном анализе, Сб. тр. Ин-та математики СО АН СССР, «Вычислительные системы», вып. 14, Новосибирск, 1964.
8. Вычислительная машина дает ответы по телефону, Электроника, **37**, № 5 (1964).
9. Загоруйко Н. Г., Об обмене устной информацией между человеком и вычислительными системами, Сб. тр. Ин-та математики СО АН СССР, «Вычислительные системы», вып. 10, Новосибирск, 1964.
10. Загоруйко Н. Г., Погрешности вычисления энергии и огибающей речевого сигнала на ЭВМ, Сб. тр. Ин-та математики СО АН СССР, «Вычислительные системы», вып. 10, Новосибирск, 1964.
11. Загоруйко Н. Г., Волошин Г. Я., Елкина В. Н., Автоматическое опознавание звуковых образцов, Сб. тр. Ин-та математики СО АН СССР, «Вычислительные системы», вып. 14, Новосибирск, 1964.
12. Какаридзе А. Г., Некоторые вопросы кодирования гласных звуков речи, Тр. Ин-та электроники, автоматики и телемеханики АН ГрузССР, **1** (1960).
13. Какаридзе А. Г., Экспериментальное устройство для автоматического различения ограниченного набора речевых команд, Элементы вычислительной техники и машинный перевод, Ин-т электроники, автоматики и телемеханики АН ГрузССР, Тбилиси, 1964.
14. Колдуэлл В., Глэссер Э., Стюарт Д., Аналоговая модель уха, Сб. Проблемы бионики, М., 1965.
15. Лейбман Ю. А., Соболев В. Н., Преобразователь аналог-цифра для ввода речевого сигнала в вычислительную машину, Электросвязь, № 8 (1963).
16. Мясников Л. Л., Объективное распознавание звуков речи, Ж. технич. физ., **13**, № 3 (1943).
17. Мясников Л. Л., Звуки речи и их объективное распознавание, Вестн. Ленингр. ун-та, № 3 (1946).
18. Мясников Л. Л., Физические исследования звуков русской речи, Изв. АН СССР. Сер. физ., **13**, № 6 (1949).
19. Мюллер П., Мартин Т., Путцрат Ф., Общие принципы операций в нейронных сетях и их приложение к распознаванию акустических образов, Сб. Проблемы бионики, М., 1965.
20. Пирогов А. А., Теоретические соображения о способе кодирования и синтеза речевой информации с помощью гармонической функции, Докл. на Всесоюзн. совещ. секции речи Комиссии по акустике АН СССР, М., 1958.
21. Пирогов А. А., Гармоническая система сжатия спектров речи, Электросвязь, № 3 (1959).
22. Сапожков М. А., Речевой сигнал в кибернетике и связи, М., 1963.
23. Солодовников В. В., Статистическая динамика линейных систем автоматического управления, М., 1960.
24. Фант Г., Акустическая теория речеобразования, М., 1964.
25. Харкевич А. А., Спектры и анализ, М., 1953.
26. Харкевич А. А., Очерки общей теории связи, М., 1955.
27. Храповицкий А. В., Теоретическое и экспериментальное исследование косинусо-логарифмического вокодера, Докл. на Всесоюзн. совещ. секции речи Комиссии по акустике АН СССР, М., 1959.
28. Цемель Г. И., К определению инвариантных признаков смычных (взрывных) звуков по сигналам клиппированной речи, Изв. АН СССР. ОТН. Энергетика и автоматика, № 4 (1959).
29. Цемель Г. И., Автоматическое распознавание звуков речи, Зарубежная радиоэлектроника, № 4 (1962).
30. Цемель Г. И., Опознание небольшого набора слов по характерным признакам речевого сигнала, Сб. Проблемы передачи информации, вып. 16, М., 1964.

31. Чистович Л. А., Влияние частотных ограничений на разборчивость русских согласных звуков, Докл. на Всесоюзн. совещ. секции речи Комиссии по акустике АН СССР, М., 1956.
32. Abstracts of Papers on Speech Analysis, Stockholm Speech Communications Seminar, 1962, Royal Institute on Technology, Stockholm, Sweden, The Journal of the Acoustical Society of America (=JASA), **35**, No. 7 (1963).
33. Bennett W. R., The Correlatograph. A Machine for Continuous Display of Short Term Correlation, Bell System Techn. J., **32**, Sept. 1953.
34. Billings A. R., Simple Multiplex Vocoder, Electronic and Radio Engr, **36**, No. 5 (1959).
35. Billings A. R., Communication Efficiency of Vocoders Comparison of Low-Power and Conventional Systems, Electronic and Radio Engr, **36**, No. 12 (1959).
36. Bogert B. P., Vobanc — a Two-to-One Speech Band-Width Reduction System, JASA, **28**, No. 3 (1956).
37. Borenstein D. P., Spectral Characteristics of Digit-Simulating Speech Sounds, Bell System Techn. J., **42**, No. 6 (1963).
38. Brick D. B., Pick G. G., Microsecond Word-Recognition System, IEEE Trans. Electronic Comput. **EC-13**, No. 1 (1964).
39. Campanella S. I., A Survey of Speech Bandwidth Compression Techniques, IRE Trans. Audio, **AU-6**, No. 5 (1958).
40. Campanella S. I., Coulter D. C., Engler P., Speech Recognition by Formant Pattern Matching in N-dimensional Space, JASA, **36**, No. 5 (1964).
41. Campanella S. I., Coulter D. C., Irons R., Influence of Transmission Error on Formant Coded Compressed Speech Signals, J. Audio Engng Soc., **10**, No. 2 (1962).
42. Carré R., Lancia R., Paillé J., Gsell R., Étude et réalisation d'un détecteur de mélodie pour analyse de la parole, Onde électr., **43**, No. 434 (1963).
43. Chang S. H., Two Schemes of Speech Compression System, JASA, **28**, No. 4 (1956).
44. Chang S. H., Pihl C. E., Wiren I., The Intervalgram as a Visual Representation of Speech Sounds, JASA, **23**, No. 6 (1951).
45. Cherry C., Halle M., Jacobson R., Toward a Logical Description of Languages in Their Phonemic Aspects, Language, **29**, No. 1 (1953).
46. Clapper C. L., Digital Circuit Techniques for Speech Analysis, IEEE Trans. Communication and Electronics, **CE-11**, No. 66 (1963).
47. Cramer B., Sprachsynthese zur Übertragung mit sehr geringer Kanalkapazität, Nachrichtentechn. Z., H. 8 (1964).
48. David E. E., Mc Donald H. S., Note on Pitch-Synchronous Processing of Speech, JASA, **28**, No. 6 (1956).
49. David J., Schroeder M. K., Logan B. F., Prestigiacomo H. J., Voice-Excited Vocoders for Practical Speech Bandwidth Reduction IRE Trans. Inform. Theory, **IT-8**, No. 5 (1962).
50. Davis K. H., Bidulph R., Balashek S., Automatic Recognition of Spoken Digits, JASA, **24**, No. 6 (1952).
51. Denes P., The Design and Operation of the Mechanical Speech Recognizer at University College London, J. Brit. IRE, **19**, No. 4 (1959).
52. Denes P., Mathews M. V., Spoken Digit Recognition Using Time-Frequency Pattern Matching, JASA, **32**, No. 11 (1960).
53. Deuber G., Varied Approaches Used to Develop System for Reliable Recognition of Voice Commands, Electronic News, **8**, No. 383 (1963).
54. Deweze A., Techniques des reconnaissance automatique des formes visuélles et sonores, Automatisme, **9**, No. 3 (1964).
55. Dersch W. C., Speech Operated Safety Switch, Electronics, **36**, No. 25 (1963).
56. Dolansky L. O., Instantaneous Pitch-Period Indicator, JASA, **27**, No. 1 (1955).
57. Dreyfus-Graf J., Phonétographe et Subformants, Bull. Techn. PTT, Bern, n° 2 (1957).
58. Dreyfus-Graf J., Phonétographe: Présent et Future, Bull. Techn. PTT, Bern, n° 5 (1961).
59. Dudley H. W., Remaking Speech, JASA, **11**, No. 2 (1939).
60. Dudley H. W., Speech Analysis and Synthesis System, JASA, **22**, No. 6 (1950).
61. Dudley H. W., Phonetic Pattern Recognition for Narrowband Transmission, JASA, **30**, No. 8 (1958).
62. Dunn H. K., Methods of Measuring Vowel Formant Bandwidths, JASA, **33**, No. 12 (1961).
63. Fano R. M., The Information Theory Point of View in Speech Communication, JASA, **22**, No. 10 (1950).
64. Fant G., Acoustic Theory of Speech Production with Calculations Based on X-Ray Studies of Russian Articulations, Mouton & Co, s'Gravenhage, 1960.

65. Fant G., Fintoft K., Liliencrants J., Lindblom B., Màrtony J., Formant-Amplitude Measurements, JASA, **35**, No. 11 (1963).
66. Flanagan I. L., A Difference Limen for Vowel Formant Frequency, JASA, **27**, No. 3 (1955).
67. Flanagan J. L., Automatic Extraction of Formant Frequences from Continuous Speech, JASA, **28**, No. 1 (1956).
68. Flanagan J. L., Band-Width and Channel Capacity Necessary to Transmit the Formant Information of Speech, JASA, **28**, No. 4 (1956).
69. Flanagan J. L., A Resonance-Vocoder and Baseband Complement: A Hybrid System for Speech Transmission, IRE Trans. Audio, **AU-8**, May—June **1960**.
70. Forgie J. W., Hughes C. W., A Real-Time Speech Input System for a Digital Computer, JASA, **30**, No. 7 (1958).
71. Forgie J. W., Forgie C. D., Results Obtained from a Vowel Recognition Computer Programme, JASA, **31**, No. 9 (1959).
72. Fujisaki H., Automatic Extraction of Fundamental Period of Speech by Auto-Correlation Analysis and Peak Detaction, JASA, **32**, No. 11 (1960).
73. Gabor D., New Possibilities in Speech Transmission, J. IRE, **94**, Nov. 1948.
74. Galdwell W. F., Recognition of Sounds by Cochlear Patterns, IEEE Trans. Military Electronics, **MIL-7**, No. 2—3 (1963).
75. Gerstman L. J., Liberman A. M., Delattre P. C., Cooper F. S., Rate and Duration of Change in Formant Frequency as Cues for Identification of Speech Sounds, JASA, **26**, No. 9 (1954).
76. Gold B., Rader C., Bandpass Compressor: A New Type of Speech-Compression Device, JASA, **36**, No. 6 (1964).
77. Golden R. M., MacLean D. I., Prestigiacomo A. I., Frequency Multiplex System for a 10-Spectrum-Channel Voice-Excited Vocoder, JASA, **36**, No. 10 (1964).
78. Gribenski A., The Pitch of Sound, Its Measuring and Perception, Nature, **173**, No. 4 (1957).
79. Haggard M. P., In Defence of the Formant, Phonetica, **10**, No. 3—4 (1963).
80. Harris C. M., Waite W. M., Gaussian-Filter Spectrum Analyzer, **35**, No. 4 (1963).
81. Harris C. M., Weiss M. R., Pitch Extraction by Computer Processing of High-Resolution Fourier Analysis Data, JASA, **35**, No. 3 (1963).
82. Hellwarth G. A., Speech-Formant Measurement with a Continuously Tuned Automatic Tracking Filters, JASA, **35**, No. 5 (1963).
83. Heydeman P., Ein Modellversuch zum Frequenzunterscheidungsvermögen des Ohres, Acustica, **13**, No. 2 (1963).
84. Hillix W., Use of Two Nonacoustic Measures in Computer Recognition of Spoken Digits, JASA, **35**, No. 12 (1963).
85. Howard C. R., Speech Analysis Synthesis Scheme Using Continuous Parameter, JASA, **28**, No. 6 (1956).
86. Howard C. R., Chang S. H., Carrabes M. J., Analysis and Synthesis of Formants and Moments of Speech Spectra, JASA, **28**, No. 4 (1956).
87. Huggins W. H., A Phase Principle for Complex-Frequency Analysis and Its Implications in Auditory Theory, JASA, **24**, No. 6 (1952).
88. Hughes G. W., Identification of Speech Sounds by Means of a Digital Computer, JASA, **31**, No. 1 (1959).
89. Husson R., Zur Spektralstruktur menschlicher Vokale aller Stimmstärken, Phonetica, No. 1—2 (1963).
90. Inomata S., An Auditory Pattern Processing Model, IEEE Trans. Inform. Theory, **IT-9**, No. 4 (1963).
91. Inomata S., Speech Recognition and Generation by a Digital Computer, Res. Electrotechn. Labs, No. 645 (1963).
92. Ithell A. H., A Determination of the Acoustical Input Impedance Characteristics of Human Ears, Acustica, **13**, No. 4 (1963).
93. Jaffe J., Cassotta L., Feldstein S., Markovian Model of Time Patterns of Speech, Science, **144**, No. 3260 (1964).
94. Jakobson R., Fant G. G., Halle M., Preliminaries to Speech Analysis, The Distinctive Features and Their Correlates, Massachusetts Institute of Technology, Acoust. Lab. Techn. Rept, No. 13 (1952).
95. Jakobson R., Die Verteilung der stimmhaften und stimmlosen Geräuschlaute im Russischen, Festschrift für Max Vasmer, Berlin, 1956.
96. Johnson W., System to Generate Speech from Written Pattern Shown, Electronic News, **8**, No. 392 (1963).
97. Kacprowski J., Speech Compression by Means of Analysis-Synthesis Methods. Polish Academy of Sciences, Proc. Vibration Probl., **5**, No. 3 (1964).

98. Kadokava J., Nakata K., Formant Frequency Extraction by Analysis-by-Synthesis Technique, J. Radio Res. Labs, **10**, No. 49 (1963).
99. Kadokawa J., Nakata K., Analysis of Speech by Vocal Tract Configuration, J. Radio Res. Labs, **11**, No. 54 (1964).
100. Kalfaian M. V., Phonetic Typewriter of Speech, JASA, **36**, No. 6 (1964).
101. Karplus H. B., Correlation Hypothesis to Explain the Fine Frequency Discrimination of the Ear, JASA, **35**, No. 5 (1963).
102. Klass P. I., Vocoder Increases Channels Security, Aviat. Week, **73**, No. 4 (1960).
103. Klumpp R. G., Webster I. C., Intelligibility of Time-Compressed Speech, JASA, **33**, No. 3 (1961).
104. Koenig W., A New Frequency Scale for Acoustic Measurements, Bell Labs, Rec., **27**, No. 7 (1949).
105. Ladefoged P., Acoustic Correlate of Subglottal Activity, JASA, **35**, No. 5 (1963).
106. Ladefoged P., McKinney N. P., Loudness, Sound Pressure and Subglottal Pressure in Speech, JASA, **35**, No. 4 (1963).
107. Latil de, P., La parole est aux calculateurs, Électronique Industr., n° 76 (1964).
108. Lehiste I., Peterson G. E., Some Basic Considerations in the Analysis of Intonation, JASA, **33**, No. 4 (1961).
109. Lieberman P., Perturbations in Vocal Pitch, JASA, **33**, No. 5 (1961).
110. Liberman A. M., Ingeman F., Lisker L., Delattre P. C., Cooper F. S., Minimal Rules for Synthesizing Speech, JASA, **31**, No. 11 (1959).
111. Licklider I. C., Effects of Amplitude Distortion Upon the Intelligibility of Speech, JASA, **18**, No. 2 (1964).
112. Licklider I. C., The Intelligibility of Amplitude-Dichotomized, Time-Quantized Speech Waves, JASA, **22**, No. 6 (1950).
113. Licklider I. C., Influence of Phase Coherence upon the Pitch of Complex Periodic Sounds, JASA, **27**, No. 5 (1955).
114. Licklider I. C., Man-Computer Symbiosis, IRE Trans. **HFE-1**, No. 1 (1960).
115. Machines Controlled by Spoken Commands, Datamation, **8**, No. 6 (1962).
116. Martin T. B., Talvage I. I., Application of Neural Logic to Speech Analysis and Recognition, IEEE Trans. Military Electronics, **MIL-7**, No. 2—3 (1963).
117. Miller I. E., Mathews M. V., Investigation of the Glottal Waveshape by Automatic Inverse Filtering, JASA, **35**, No. 11 (1963).
118. Miller R. L., Improvements in the Vocoder, JASA, **25**, No. 4 (1953).
119. Nicolau E., Weber I., Gavăt S., Aparate pentru recunoasterca automata ă vocalelor, Automatica çi electronica, **7**, No. 6 (1963).
120. Noll A. M., Short-Time Spectrum and «Cepstrum» Techniques for Vocal-Pitch Detection, JASA, **36**, No. 2 (1964).
121. Oeken F. W., Some Critical Observation on the Formant Theory of Vowel Recognition, Phonetica, **10**, No. 1—2 (1963).
122. Olson H. F., Belar H., Phonetic Typewriter, JASA, **28**, No. 6 (1956).
123. Olson H. F., Belar H., Phonetic Typewriter III, JASA, **33**, No. 11 (1961).
124. Olson H. F., Belar H., Syllable Analyzer, Coder and Synthesizer for the Transmission of Speech. IRE Trans. Audio, **AU-10**, No. 11 (1962).
125. Olson H. F., Belar H., Sobrino R., Demonstration of a Speech Processing System Consisting of a Speech Analyzer, Translator, Typer and Synthesizer, JASA, **34**, No. 10 (1962).
126. Olson H. F., Belar H., Performance and a Code-Operated Speech Synthesizer, JASA, **36**, No. 5 (1964).
127. Paul A. P., House A. S., Stevens K. N., Automatic Reduction of Vowel Spectra: An Analysis-by-Synthesis Method and Its Evaluation, JASA, **36**, No. 2 (1964).
128. Peterson G. E., Design of Visible Speech Devices, JASA, **26**, No. 3 (1954).
129. Peterson E., Cooper F. S., Peakpicker: A Band-Width Compression Device, JASA, **29**, No. 6 (1957).
130. Peterson G. E., Lehiste I., Identification of Filtered Vowels, JASA, **31**, No. 6 (1959).
131. Peterson G. E., Sivertsen E., Subrahmanyam D. L., Intelligibility of Diphasic Speech, JASA, **28**, No. 3 (1956).
132. Petrick S. R., Talking to a Computer, New Scientist, No. 235 (1961).
133. Phyfe D. L., Toffler I. E., Some Features of the Army Channel Vocoder, JASA, **35**, No. 12 (1963).
134. Pick G. G., Gray C. B., Brick D. B., The Solenoid Array — a New Computer Element, IEEE Trans. Electronic Computers, **EC-13**, No. 1 (1964).
135. Pinson E. N., Pitch-Synchronons Time-Domain Estimation of Formant Frequencies and Bandwidth, JASA, **35**, No. 8 (1963).
136. Pollack I., Picket L., Effect of Noise and Filtering on Speech Intelligibility at High Levels, JASA, **29**, No. 12 (1957).

137. Potter R. K., Knopp G. A., Green H. C., Visible Speech, Van Nostrand, New York, 1947.
138. Potter R. K., Steinberg J. C., Toward the Specification of Speech, JASA, **22**, No. 6 (1950).
139. Proceedings of the Speech Communication Seminar, Stockholm, Aug. 29 to Sept. 1, 1962, Publ. Speech Transmission Lab. Royal Inst. Technology, Stockholm, 1964.
140. Pruzansky S., Bricner P. D., Automatic Talker Recognition Using Time-Frequency Pattern Matching, JASA, **33**, No. 6 (1961).
141. Pun L., The Phonetograph, Control, **7**, January 1963.
142. Rader C., Spectra of Vocoder-Channel Signals, JASA, **35**, No. 5 (1963).
143. Rader C. M., Vector Pitch Detection, JASA, **36**, No. 10 (1964).
144. Rawley I. R., Converting Digital Data to Voice, Electronic Industr., **23**, No. 4 (1964).
145. Rais A., Vowel Recognition in Clipped Speech, Nature, **181**, No. 3 (1958).
146. Righini G. U., A Pitch Extractor of the Voice. Acustica, **13**, No. 4 (1963).
147. Rosen G., Dynamic Analog Speech Synthesizer, JASA, **30**, No. 3 (1958).
148. Sackschewsky V. E., Oestreicher H. L., Pattern Recognition as a Problem in Decision Theory and an Application to Speech Recognition, IEEE Trans. Military Electronics, **MIL-7**, No. 2—3 (1963).
149. Sakai T., Dochita S., Nagata K. I., Phonetic Typewriter, JASA, **35**, No. 7 (1963).
150. Sakai T., Inone S. I., New Instruments and Methods for Speech Analysis, JASA, **32**, No. 4 (1960).
151. Schief R., Koinzidenz-Filter als Modell für das menschliche Tonhöhenunterscheidungsvermögen, Kybernetik, **2**, H. L (1963).
152. Scholtz P. N., Bakis R., Spoken Digit Recognition Using Vowelconsonants Segmentation, JASA, **34**, No. 1 (1962).
153. Schröder M. R., New Approach to Time Domain Analysis and Synthesis, JASA, **31**, No. 6 (1959).
154. Schröder M. R., Crystal T. H., Auto-Correlation Vocoder, JASA, **32**, No. 7 (1960).
155. Schröder M. R., David E. E., A Vocoder for Transmitting 10 kc/s Speech Over a 3,5 kc/s Channel, Acustica, **10**, No. 1 (1960).
156. Seki H., A New Method of Speech Transmission by Frequency Demultiplication and Multiplication, J. Acoust. Soc. Japan, **14**, June 1958.
157. Siedler G., Untersuchungen über die Bedeutung bestimmter Tonfrequenzbänder für die Verständlichkeit synthetischer Sprache und über Änderung der Sprachverständlichkeit bei Kanalvertauschungen, Z. angew. Phys., **13**, Nr. 6 (1961).
158. Simmons P. L., Automation of Speech, Speech Synthesis and Synthetic Speech. A Bibliographical Survey from 1950—1960. IRE Trans. Audio, **AU-9**, November—December 1961.
159. Slaymaker F. H., Bandwidth Compression by Means of Vocoder. IRE Trans. Audio, **AU-9**, January—February 1960.
160. Smith C. R., A Phoneme Detector, JASA, **23**, No. 4 (1951).
161. Smith C. R., The Analysis and Automatic Recognition of Speech Sounds, Electronic Engng, **24**, No. 8 (1962).
162. Smith S. L., Man-Computer Information Transfer, Electro-Technology, **72**, August 1963.
163. Speech Recognition Gets Push From Synthesizer, Electronics, **34**, 16 (1961).
164. Steele K. W., Cassel L. E., Quality Improvement in the Channel Vocoder, JASA, **35**, No. 5 (1963).
165. Stevens K. N., Acoustical Analysis of Speech, JASA, **30**, No. 7 (1958).
166. Sund H., A Sound Spectrometer for Speech Analysis, Trans. Royal Inst. Technology, Stockholm, No. 112 (1957).
167. Suzuki I., Kadokawa J., Nakata K., Formant-Frequency Extraction by the Method of Moment Calculations, JASA, **35**, No. 9 (1963).
168. Suzuki I., Nakata K., Phonemic Classification and Recognition of Japanese Monosyllables, J. Radio Res. Labs, **10**, No. 49 (1963).
169. Suzuki I., Nakata K., Maezono K., Speech Data Analysed by Computer Program, Classification and Recognition of Japanese Monosyllables, IEEE Trans. Military Electronics, **MIL-7**, No. 2—3 (1963).
170. Tarnóczy T. H., Vowel Formant Bandwidths and Synthetic Vowels, JASA, **34**, No. 6 (1962).
171. Thierney J., Gold B., Sferrino V., Dumanian I. A., Aho E., Channel Vocoder with Digital Pitch Extractor, JASA, **36**, No. 10 (1964).
172. Toffler I. E., Wade F. B., Pitch Extractor, Using Clippers, JASA, **36**, No. 5 (1964).
173. Uhr L., Recognition of Letters, Pictures and Speech by a Discovery and Learning Program, WESCON Techn. Papers, **8**, p. 4, August 25—28, 1964.

174. Vilbig F., An Analysis of Clipped Speech, JASA, **27**, No. 1 (1955).
175. Vilbig F., Speech Compression, JASA, **28**, No. 1 (1956).
176. Vilbig F., Improvement of Simplification of the Scanvocoder and Its Connection to a Correlation Pulse Code System, JASA, **28**, No. 4 (1956).
177. Vilbig F., Haase K. H., Über einige Systeme zur Sprachbandkompression, Nachrichtentechn. Fachber., NTF, **3** (1956).
178. Vilbig F., Haase K. H., Some Systems for Speech-Band Compression, JASA, **28**, No. 4 (1956).
179. Wallace J. C., Comparative Evaluation of Pitch-Signal Indicators, JASA, **35**, No. 5 (1963).
180. Waller R., Self-programming Pattern Recognizer, Measurements and Control, No. 3 (1964).
181. Waller R., Sceptron, J. Scient. Instruments, **41**, No. 5 (1964).
182. Weiss M. R., Harris C. M., Computer Technique for High-Speech Extraction of Speech Parameters, JASA, **35**, No. 2 (1963).
183. Widrow B., Groner G. F., Hu M. I. C., Smith F. W., Specht D. F., Talbert L. R., Practical Applications for Adaptive Data-Processing Systems, WESCON Techn. Papers, **7**, No. 7 (1963).
184. Winckel F., Tonhöhenextractor für Sprache mit Gleichstromanzeige, Phonetica, Nr. 3—4 (1964).
185. Wiren I., Stubbs H. L., Electronic Binary Selection System for Phoneme Classification, JASA, **28**, No. 6 (1956).
186. Wood D. E., Hewitt T. L., New Instrumentation for Making Spectrographic Pictures of Speech, JASA, **35**, No. 8 (1963).
187. Wood D. E., New Display Formant and a Flexible-Time Integration for Spectral-Analysis Instrumentation, JASA, **36**, No. 4 (1964).
188. W. S., Dr., Identification des personnes par la spéctrographie vocale, Automatisme, **8**, n° 7—8 (1963).
189. Yoshima T., Japan Firm Builds Spoken Voice Digit Recognizer, Electronic News, **8**. No. 390 (1963).
190. Zwicker E., Möglichkeiten zur Spracherkennung über den Tastsinn mit Hilfe eines Funktionsmodells des Gehörs, Elektronische Rechenanlagen, **6**, H. 6 (1964).

Институт кибернетики
Академии наук Эстонской ССР

Поступила в редакцию
14/VI 1965

E. KÜNNAP

SUULISED KÄSUD JUHTIMISSÜSTEEMIDES

Küberneetilised masinad, mis teeksid võimalikuks vahetu suhtlemise elava ja elutu (*resp.* inimese ja masina) vahel, paneksid ühtlasi aluse uutele lülidele juhtimissüsteemides. Artiklis tutvustatakse uurimistöid automaatse kõne äratundmise valdkonnas ja kirjeldatakse nende resultaatide rakendamist.

E. KÜNNAP

SPEECH COMMANDS IN CONTROL SYSTEMS

Cybernetic machines where the speech sounds of men could be directly received by a lifeless object give possibilities to design new links for control systems. A review of researches in the region of the automatic recognition of speech sounds and some aspects of an application of their results are presented.