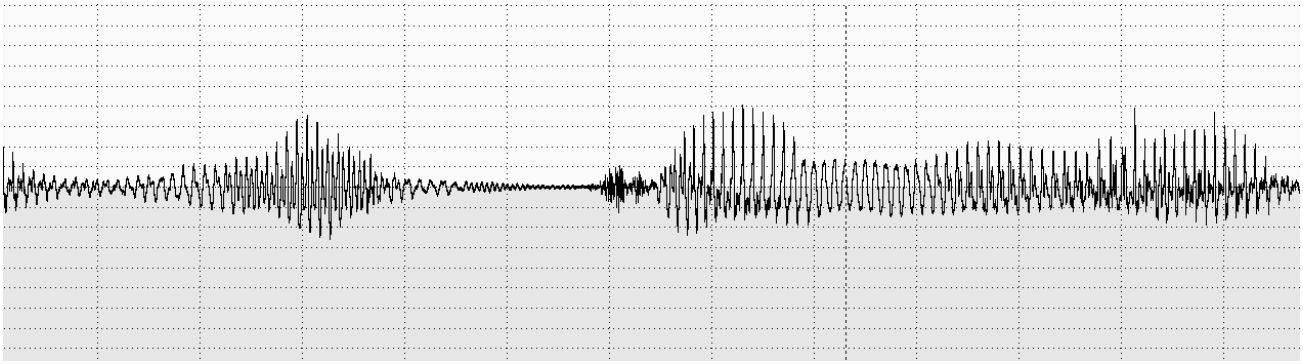
УНИВЕРСИТЕТ ИТМО

И.Б. Тампель, А.А. Карпов

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ

Учебное пособие



Санкт-Петербург

2017

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

И.Б. Тампель, А.А. Карпов

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ

Учебное пособие

РЕКОМЕДОВАНО К ИСПОЛЬЗОВАНИЮ В УНИВЕРСИТЕТЕ ИТМО
по направлению подготовки «информационные системы и технологии»
в качестве учебного пособия для реализации основных образовательных программ высшего образования магистратуры

УНИВЕРСИТЕТ ИТМО

Санкт-Петербург

2017

Тампель И.Б., Карпов А.А. АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ. Учебное пособие. – СПб: Университет ИТМО, 2017. – 152 с.

В учебном пособии рассматриваются методы автоматического распознавания речи. Материал пособия разбит на 16 разделов. Первые два раздела посвящены вопросам речеобразования и восприятия слуховой системой. В каждом разделе приведены краткие теоретические и/или практические сведения.

Пособие может быть использовано при подготовке магистров по направлению 09.04.02 ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ и аспирантов.

Рецензент: к.т.н. Кореневский М.Л.

Рекомендовано к печати Ученым советом факультета Информационных технологий и программирования 21.10.2017 г., протокол № 10

УНИВЕРСИТЕТ ИТМО

Университет ИТМО – ведущий вуз России в области информационных и фотонных технологий, один из немногих российских вузов, получивших в 2009 году статус национального исследовательского университета. С 2013 года Университет ИТМО – участник программы повышения конкурентоспособности российских университетов среди ведущих мировых научно-образовательных центров, известной как проект «5 в 100». Цель Университета ИТМО – становление исследовательского университета мирового уровня, предпринимательского по типу, ориентированного на интернационализацию всех направлений деятельности.

Содержание

Введение

Автоматическое распознавание речи является динамично развивающимся направлением в области искусственного интеллекта. За последние полвека в данной области достигнуты значительные успехи – имеется множество коммерческих приложений, которые делают вложения в данную область оправданными и выгодными. Среди таких приложений, в первую очередь, можно отметить внедрение call-центров или IVR-систем (Interactive Voice Response) – систем автоматического доступа к информации, минуя оператора. В современных call-центрах вопросы формулируются пользователем на естественном языке, и ответ синтезируется компьютером также на языке пользователя. Внедрение call-центров позволило высвободить огромное количество операторов и улучшить качество обслуживания во многих аэропортах и на железнодорожных вокзалах.

Системы автоматического распознавания речи широко применяются в медицинских исследованиях, требующих ввода информации, когда руки оператора заняты (рентгеновские), или когда требуется управлять автономными аппаратами исследования внутренних органов. Даже заполнение медицинских карт средним персоналом в продвинутых медицинских учреждениях ведётся голосом.

Важной областью применения систем автоматического распознавания и синтеза речи является помощь людям с инвалидностью, как с проблемами опорно-двигательного аппарата, так и слабовидящим (ассистивные технологии).

Следует отметить, что в России медицинские приложения систем автоматического распознавания речи практически не реализованы, что оставляет огромное поле деятельности для разработчиков.

Несмотря на значительные успехи, главная цель исследований, которая изначально подразумевалась – свободное общение человека и «машины» – пока не достигнута. Развитие направления выявило новые трудности, бросающие вызов исследователям на современном этапе, когда задача распознавания речи смыкается с проблемой понимания смысла сообщения и требует привлечения научной психологии.

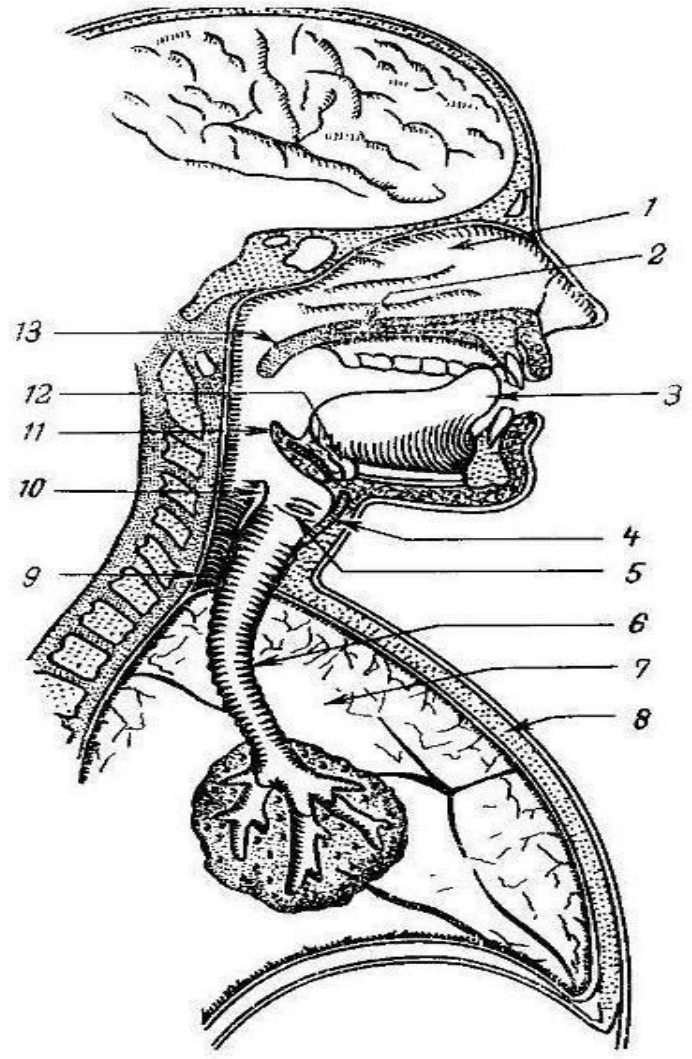
Первые два раздела пособия посвящены вопросам речеобразования и восприятия. Очевидно, что понимание структуры речевого сигнала и лежащих в его основе движений речеобразующих органов может помочь в решении задачи автоматического распознавания речи. В ещё большей степени это относится к пониманию вопросов, связанных с восприятием звуков вообще и речевых звуков в частности.

Очень важным вопросом, на который предстоит ответить в ходе изучения речеобразования и восприятия является вопрос о признаках, или параметрах речевого сигнала, которые содержат информацию, достаточную для распознавания речи. Очевидно, по самому своему смыслу, эти параметры должны являться следствием сознательно контролируемых движений речевых органов. Очевидно также, что выделение этих параметров должно являться главной задачей слуховой системы при распознавании речи.

1. РЕЧЕОБРАЗОВАНИЕ

1.1. Физиология речеобразования

Процесс речеобразования иллюстрируется на рис. 1.1. и 1.2. Благодаря создаваемому в лёгких давлению, поток воздуха устремляется в голосовой тракт, проходит через голосовые складки, может устремляться в носовую полость (если нёбная занавеска открыта) и выходит в открытое пространство, минуя возможные зубные и губные сужения.

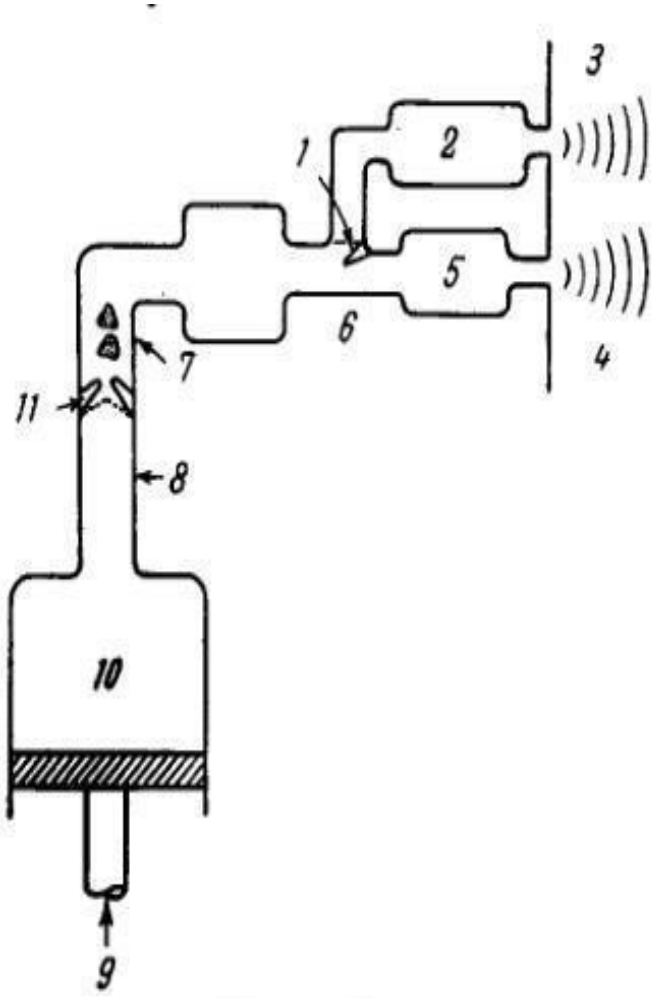


1 — носовая полость, *2* — твердое небо, *3* — язык, *4* — щитовидный хрящ, *5* — голосовые связки, *6* — трахея, *7* — легкое, *8* — грудина, *9* — пищевод, *10* — кольцеобразный хрящ, *11* — надгортанье, *12* — подъязычная кость, *13* — мягкое небо (небная занавеска)

Рис.1.1. Речевой аппарат человека [1].

Речь представляет собой звуковые колебания воздуха в диапазоне частот от 70–100 Гц до нескольких килогерц. Для того чтобы в выходящем воздушном потоке возникли колебания с такими частотами, необходимо наличие источника звука на пути воздушного потока. Источником звука могут являться:

1. Голосовые складки;
2. Турбулентный шум в сужении;
3. Шум внезапно высвободившегося воздуха при смычке (импульсный).



1 — небная занавеска, *2* — носовая полость, *3* — излучения носового тракта, *4* — излучения рта, *5* — ротовая полость, *6* — поднятая часть языка, *7* — гортанная трубка, *8* — трахея и бронхи, *9* — мускульная сила, *10* — объем легких, *11* — голосовые связки

Рис.1.2. Схематическое изображение речевого аппарата и функциональных узлов речевого тракта человека [1].

Места сужения или смычки могут быть разными для разных языков (так, в ряде языков существуют необычные для русского языка звуки, источником которых является гортанная смычка, то есть взрыв, образующийся при размыкании голосовых складок). При образовании звука /х/ и шепотной речи шумовым источником являются сведенные, но не колеблющиеся голосовые складки.

В соответствии с типом источника речевые звуки подразделяются на классы:

1. Гласные – источником звука являются только голосовые складки, проход в носовую полость перекрыт небной занавеской;

2. Щелевые (фрикативные) согласные – источником звука является турбулентный шум в сужении (глухие согласные /ф/, /с/, /ш/,..), или дополнительно голосовые складки (звонкие /в/, /з/, /ж/,..).

3. Взрывные согласные – источником звука является шум взрыва (глухие /п/, /т/, /к/), или дополнительно импульсы голосовых складок (звонкие /б/, /д/, /г/).

Кроме указанных существуют звуки, которые требуют отдельной классификации:

1. Носовые согласные. Характеризуются тем, что излучение полностью или частично осуществляется через нос. Забегая вперед, отметим, что передаточная функция голосового тракта содержит только полюса, то есть обладает только резонансами; при наличии боковой полости или параллельной ветви передаточная функция содержит также нули.

2. Русское /р/ возбуждается голосовыми складками, однако звук модулируется дрожанием кончика языка.

3. Звуки, получающиеся сочетанием рассмотренных выше (примеры на основе общеамериканского диалекта):

Полугласные /j/ you, /w/ we;

Плавные /r/ read, /l/ let.

4. Звуки, характеризующиеся динамическим характером произнесения:

дифтонги /eI/ say, /Iu/ new, /ɔI/ boy, /aU/ out, /aI/ I, /oU/ go;

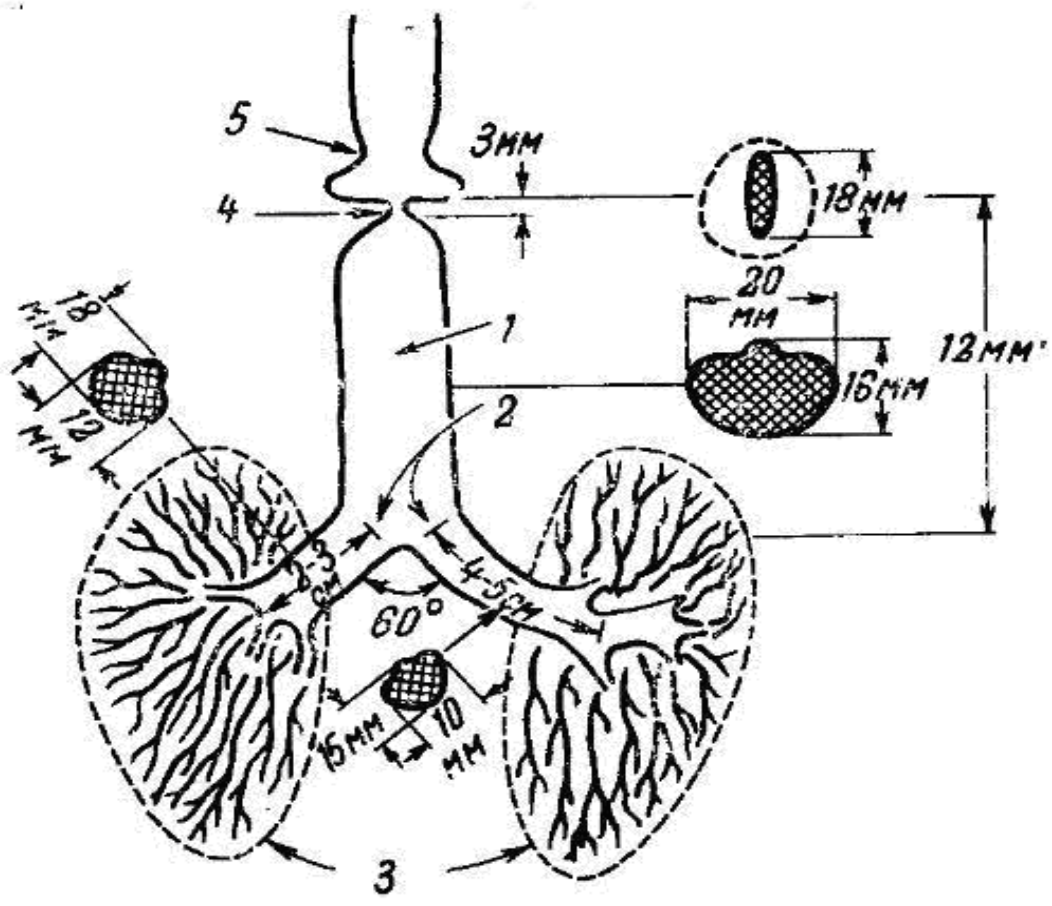
аффрикаты /tʃ/ chew, /dʒ/ jar.

1.1.1. Процесс образования звуков с голосовым возбуждением

Голосовые складки (связки) (рис.1.3) колеблются при продувании через них потока воздуха под действием эффекта Бернулли [1]. Частота колебаний голосовых складок называется основным тоном (pitch). Для полноты обзора следует упомянуть о таком этническом феномене, как двухголосое тувинское пение. В данном случае имеется два периодических источника возбуждения. Один из них – обычные голосовые складки, а второй – либо ложные складки (утолщения, расположенные над голосовыми складками), либо верхушка пищевода (известно, что люди с повреждёнными голосовыми складками овладевают «пищеводной речью»). Частота колебаний голосовых складок при обычной речи находится в пределах 60–180 Гц для мужчин, 160–350 Гц для женщин и 200–650 Гц для детей (указанные границы чисто ориентировочные).

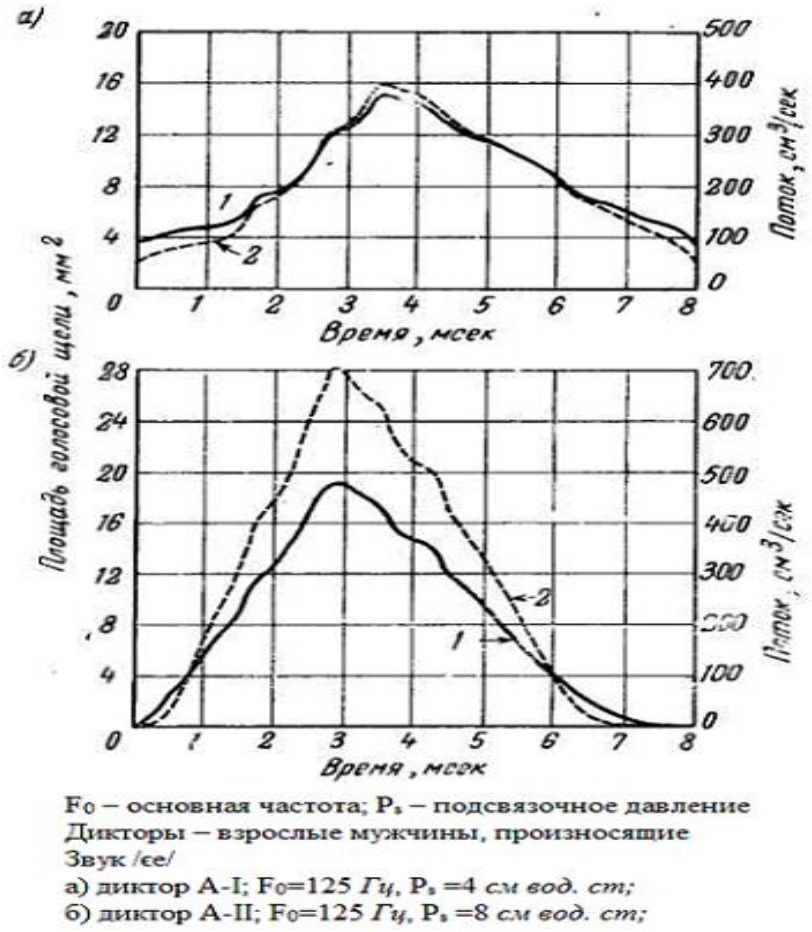
При пении частота колебаний голосовых складок может достигать 2 кГц.

Форма импульсов объемной скорости (скорость потока, умноженная на площадь сечения в данной точке – величина, сохраняющаяся вдоль всего тракта), исходящих из голосовой щели, неплохо аппроксимируется треугольником или полуволной синуса (рис. 1.4). При этом скважность может достигать 40%, то есть складки могут смыкаться на продолжительное время за счет упругости тканей при столкновении. Для подобных импульсов спектр спадает приблизительно со скоростью 12 дБ/окт. Понятно, что чем более угловатую форму имеют импульсы, и чем больше разрывов имеет производная объемной скорости, тем более длинный хвост будет у спектра. Форма импульсов голосового источника, в основном, и определяет тембр голоса и «полетность» певческого голоса.



1 - трахея, 2 - бронхи, 3 - лёгкие (макс. 4-5 л),
4 - голосовые связки, 5 - ложные связки

Рис.1.3. Схематическое изображение органов подгортанной системы человека [1].



F₀ – основная частота; Pₛ – подсвязочное давление
Дикторы – взрослые мужчины, произносящие
Звук /єе/
а) диктор А-I; F₀=125 *Гц*, Pₛ =4 *см вод. ст;*
б) диктор А-II; F₀=125 *Гц*, Pₛ =8 *см вод. ст;*

Рис.1.4. Кривые изменения площади голосовой щели (а), кривые объёмной скорости для одного периода основного тона (б) [1].

1.2. Передаточная функция голосового тракта

Передаточная функция голосового тракта рассчитывается, исходя из того, что для значимых для восприятия частот (<4000 Гц) акустическая волна с достаточной точностью является плоской. Распространение звука в этом случае описывается одномерным (зависящим от координаты вдоль оси тракта) уравнением Вебстера:

$$\frac{1}{S(x)}\frac{\partial}{\partial x}\left[S(x)\frac{\partial p}{\partial x}\right]=\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2}, \qquad (1.1)$$

где $S(x)$ – площадь поперечного сечения как функция расстояния от голосового источника по оси тракта, p – звуковое давление, c – скорость звука, t – время.

Даже в одномерном случае для голосового тракта уравнение Вебстера можно решить только численно. При этом не учитывается импеданс стенок тракта и потери энергии на границах и на трение.

Для волновода постоянного сечения ($S(x)$=const) уравнение Вебстера превращается в одномерное волновое уравнение для плоской волны в пространстве:

$$\frac{\partial^2 p}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{1.2}$$

Такое же по форме уравнение справедливо для объёмной скорости *u***.**
Давление и скорость связаны уравнениями:

$$\begin{aligned} -\frac{\partial p}{\partial x} &= \rho \frac{\partial u}{\partial t}; \\ -\frac{\partial u}{\partial x} &= \frac{1}{\rho c^2}\frac{\partial p}{\partial t}, \end{aligned} \tag{1.3}$$

где ρ – плотность воздуха.
Общее решение уравнения (1.2) имеет вид:

$$p(x,t) = \varphi(t - \frac{x}{c})\psi(t + \frac{x}{c}), \tag{1.4}$$

где ψ и φ – функции, определяемые из начальных или граничных условий.

Если сечение голосового тракта постоянно по длине, его можно представить в первом приближении как волновод, закрытый со стороны связок и открытый со стороны губ. При этом в точке $x=0$ (на связках) скорость и частная производная давления по x равны 0. В точке $x=L$ (на губах) звуковое давление и частная производная скорости по x равны 0, где L – длина тракта.

На основании (1.2), (1.3) и (1.4) будем искать скорость и звуковое давление в виде:

$$\begin{aligned} u(x,t) &= u^+(t - x/c) - u^-(t + x/c), \\ p(x,t) &= \rho c[u^+(t - x/c) - u^-(t + x/c)]. \end{aligned} \tag{1.5}$$

Найдём выражение для скорости при условии, что на закрытом конце возбуждаются колебания

$$u(t) = \sin(\omega t). \tag{1.6}$$

Будем искать решение для скорости в виде:

$$u(x,t) = a[\sin(\varphi + \omega(t - \frac{x}{c})) + \sin(\psi + \omega(t + \frac{x}{c}))]. \tag{1.7}$$

Используя формулу суммы синусов, получим:

$$u(x,t) = 2a\sin(\frac{\varphi + \psi}{2} + \omega t)\cos(\frac{\varphi - \psi}{2} - \frac{\omega x}{c}). \tag{1.8}$$

Учитывая (1.6), находим: $\varphi + \psi = 0$, или $\varphi = -\psi$, отсюда получаем:

$$u(x,t) = 2a\sin(\omega t)\cos(\varphi - \frac{\omega x}{c}). \tag{1.9}$$

Поскольку

$$\left.\frac{\partial u(x,t)}{\partial x}\right|_{x=l} = 2a\frac{\omega}{c}\sin(\omega t)\sin(\varphi - \frac{\omega l}{c}) = 0, \tag{1.10}$$

можем положить $\varphi = \omega l/c$.

Подставляя значение φ в (1.9) и снова учитывая (1.6), получим формулу для скорости:

$$u(x,t) = \frac{\sin(\omega t)\cos(\frac{\omega(l-x)}{c})}{\cos(\frac{\omega l}{c})}. \qquad (1.11)$$

Таким образом, видим, что возбуждая объёмную скорость с единичной амплитудой на входе волновода, на выходе имеем скорость с амплитудой $1/cos(\omega l/c)$. Отношение выходной скорости к входной называется передаточной функцией.

Если принять скорость звука в 350 м/с (скорость во влажном воздухе при 36^0), а длину голосового тракта 17.5 см (длина мужского тракта), то график передаточной функции будет иметь вид (рис.1.5.):

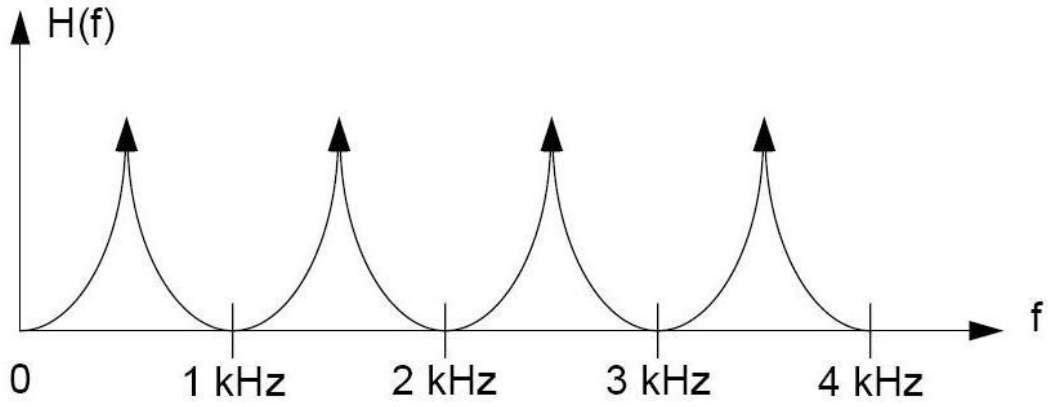


Рис.1.5. Передаточная функция волновода постоянного сечения без потерь [2].

Максимумам в спектре соответствуют стоячие волны с длиной волны:

$$\lambda(n) = \frac{l}{1/4 + n/2}, n = 0,1,..., \qquad (1.12)$$

или частотами:

$$F(n) = \frac{c}{4l}(1+2n), n = 0,1,.... \qquad (1.13)$$

Для принятых выше скорости звука и длины голосового тракта частоты будут равны: 500, 1500, 2500, 3500 Гц. Отметим, что многоточие не имеет особого практического смысла, поскольку выше 4000–5000 Гц длина волны становится сравнимой с поперечными размерами тракта и, следовательно, уравнение Вебстера не годится для описания процессов распространения звука.

Если учесть потери в среде, в голосовой щели и на стенках голосового тракта, излучение на губах, то максимумы перестанут быть бесконечными и

слегка сместятся, но общая картина качественно не изменится. Делаем важный вывод: передаточная функция голосового тракта и, следовательно, речевой сигнал характеризуются максимумами в спектре, отстоящими друг от друга на несколько сотен герц и зависящими, в основном, от формы голосового тракта. Для гласных эти максимумы называются формантами.

1.2.1. Расчёт передаточной функции с помощью электроаналогий

Голосовой тракт произвольной формы можно представить как набор цилиндрических секций. Каждую секцию можно описать как электрическую цепь:

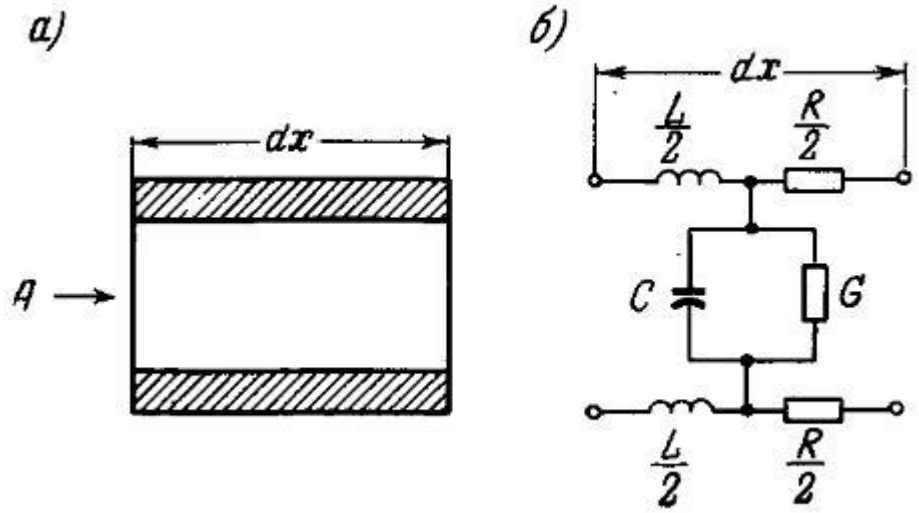


Рис.1.6. Электрический эквивалент цилиндрического отрезка трубы [1].

Аналогом звукового давления является напряжение, аналогом скорости – ток.

$$L = \frac{\rho}{S}\left(1 + \frac{S}{D}\sqrt{\frac{\mu}{4\pi f \rho}}\right),$$

где ρ – плотность воздуха, S – площадь поперечного сечения секции, D – периметр сечения секции, f – частота, μ – коэффициент вязкости;

$$R = \frac{D}{S^2}\sqrt{\pi f \rho \mu},$$

$$C = \frac{S}{\rho c^2},$$

$$G = D\frac{\eta - 1}{\rho c^2}\sqrt{\frac{\pi \lambda f}{c_p \rho}},$$ где η – адиабатическая постоянная (7/5), λ – теплопроводность воздуха, c_p – удельная теплоёмкость.

Нагрузочное излучение через рот аппроксимируется нагрузкой на круглый поршень в сфере или бесконечном плоском экране.

$$Z \approx 2\left(\frac{\pi a r}{c}\right)^2 + i\frac{16}{3}\frac{fa}{c}; \frac{2\pi f a}{c} << 1,$$

где a – радиус поршня.

Распространение звука вокруг головы рассчитывают как излучение пульсирующей сферы. Звуковое давление как функция расстояния от центра сферы r имеет вид:

$$p(r) = i\frac{2\pi f \rho a^2 u_0}{r} e^{-i\frac{2\pi f}{c}r},$$

где a – радиус сферы, u_0 – амплитуда скорости колебаний сферы.

Звуковое давление при прочих равных параметрах растёт пропорционально частоте f, то есть со скоростью 6 децибел на октаву.

Расчёты, проводимые с вышеприведёнными формулами, дают хорошее совпадение с экспериментальными данными. Расчётные и измеренные значения формант приведены в таблице 1.1.

Таблица 1.1.

Расчётные (для закрытой голосовой щели) и измеренные [3] значения формант гласных русского языка.

Гласные/ форманты	F1		F2		F3	
	(расчётн.)	(измер.)	(расчётн.)	(измер.)	(расчётн.)	(измер.)
/у/	301	300	619	625	2388	2500
/о/	549	535	859	780	2368	2500
/а/	686	700	1075	1080	2432	2600
/е/	453	440	1954	1800	2737	2550
/и/	278	240	2263	2250	2924	3200
/ы/	326	300	1477	1480	2314	2230

1.3. Турбулентный и импульсный источники звука

При образовании звуков /ш/, /ф/, /х/, /с/… источником служит турбулентный шум в сужении голосового тракта или складок. Спектр шумового источника имеет плоскую вершину, частота которой приблизительно равна $0.2\frac{v}{\sqrt{S}}$, где v – скорость потока в сужении, S – площадь сужения.

Звонкие фрикативные /ж/, /в/, /з/… образуются при одновременной работе голосового и турбулентного источников. При этом в сужении турбулентный шум модулируется импульсами голосового источника.

Взрывные согласные /п/, /б/, /к/, /г/, /т/, /д/ образуются при полной смычке в некоторой области голосового тракта, повышении давления за смычкой и резком высвобождении воздуха в результате открытия смычки. Взрыв сопровождается фрикативным шумом.

1.4. Носовые согласные

При опускании нёбной занавески звук проникает в носовую полость и излучается через нос. Голосовой тракт может быть полностью или частично закрыт на губах. Так формируются носовые (назальные) согласные звуки. Расчёт методом электроаналогий показывает, что в этом случае основной особенностью передаточной функции является ноль или антиформанта на частоте около 2 кГц.

1.5. Выводы

Таким образом, речевой сигнал представляет собой квазипериодический сигнал для вокализованных и случайный шум в области 3–6 кГц для шумовых и взрывных звуков. Спектральные максимумы чередуются через 300–700 Гц по частоте и через 150–200мс по времени (силлабическая частота).

В качестве иллюстрации на рис. 1.7. приведены огибающая и «видимая речь» (изображение спектра на плоскости, где по оси абсцисс отложено время, по оси ординат – частота, а амплитуда отображается степенью зачернения или цветом) продолжительных высказываний для широкополосного (окно=6мс) и узкополосного (окно=30мс) анализа Фурье. Заметим, что горизонтальные полосы на узкополосном спектре представляют собой гармоники основного тона. На широкополосном спектре они сливаются из-за невысокого разрешения по частоте, однако форманты при этом выделяются лучше.

Зададимся, наконец, вопросом, ради которого приводились все эти факты: какие параметры речевого сигнала являются существенными для передачи смысла сообщения, то есть являются характеристикой данного языка, а какие представляют индивидуальные особенности говорящего? (Этот же вопрос в несколько другой плоскости будет задан в разделе «Слуховая система».)

Данные психоакустики позволяют считать установленным, что существенными для восприятия являются значения энергии колебаний в довольно широких спектральных зонах. Так, для восприятия гласных существенны максимумы огибающей спектра, как правило, совпадающие с формантами. Считается, что для идентификации гласной достаточно двух первых формант, поэтому гласные часто изображают на плоскости *F1, F2*, где они образуют вытянутый треугольник. Возможно, третья форманта является дополнительным, избыточным признаком гласной. Некоторым подтверждением этой точки зрения является факт, что человек может научиться читать видимую речь (см. роман А.Солженицына «В круге первом»). Заранее отметим, что это умение никак не отразилось на создании систем автоматического распознавания речи. Либо виртуозы чтения не могут вербализовать правила, которыми они пользуются при чтении, то есть используются столь же мало познанные процессы зрительного восприятия, либо ошибки распознавания по видимой речи больше, чем в системах автоматического распознавания (уровень ошибок распознавания по видимой речи никто не измерял).

На рис. 1.8. приведены примеры видимой речи для различных слогов. Дополнительно приведены число переходов огибающей через ноль, полная

энергия и низкочастотная энергия (125–750 Гц). На основании этих рисунков можно получить представление о характере спектров для отдельных звуков речи и их сочетаний.

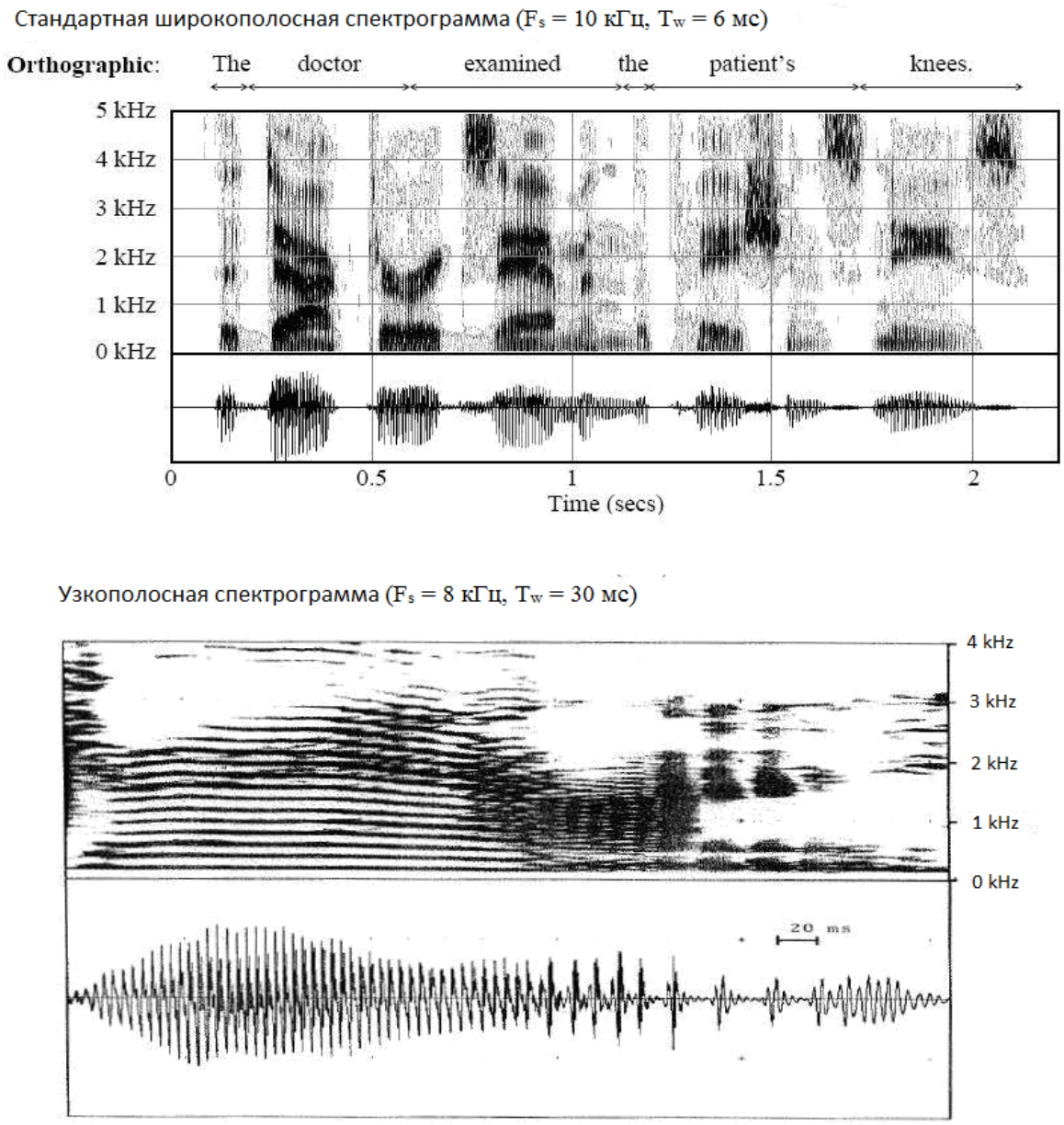
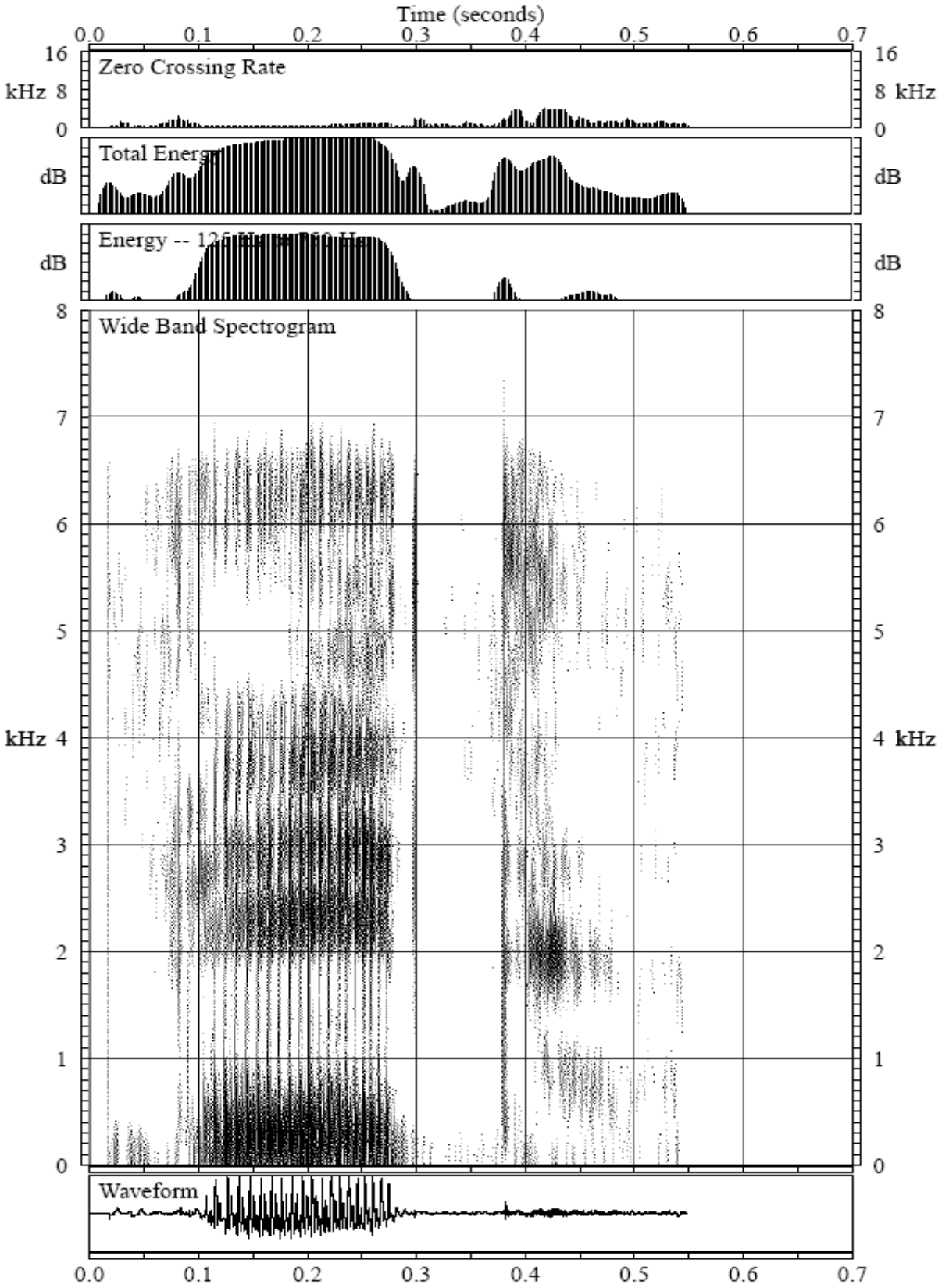
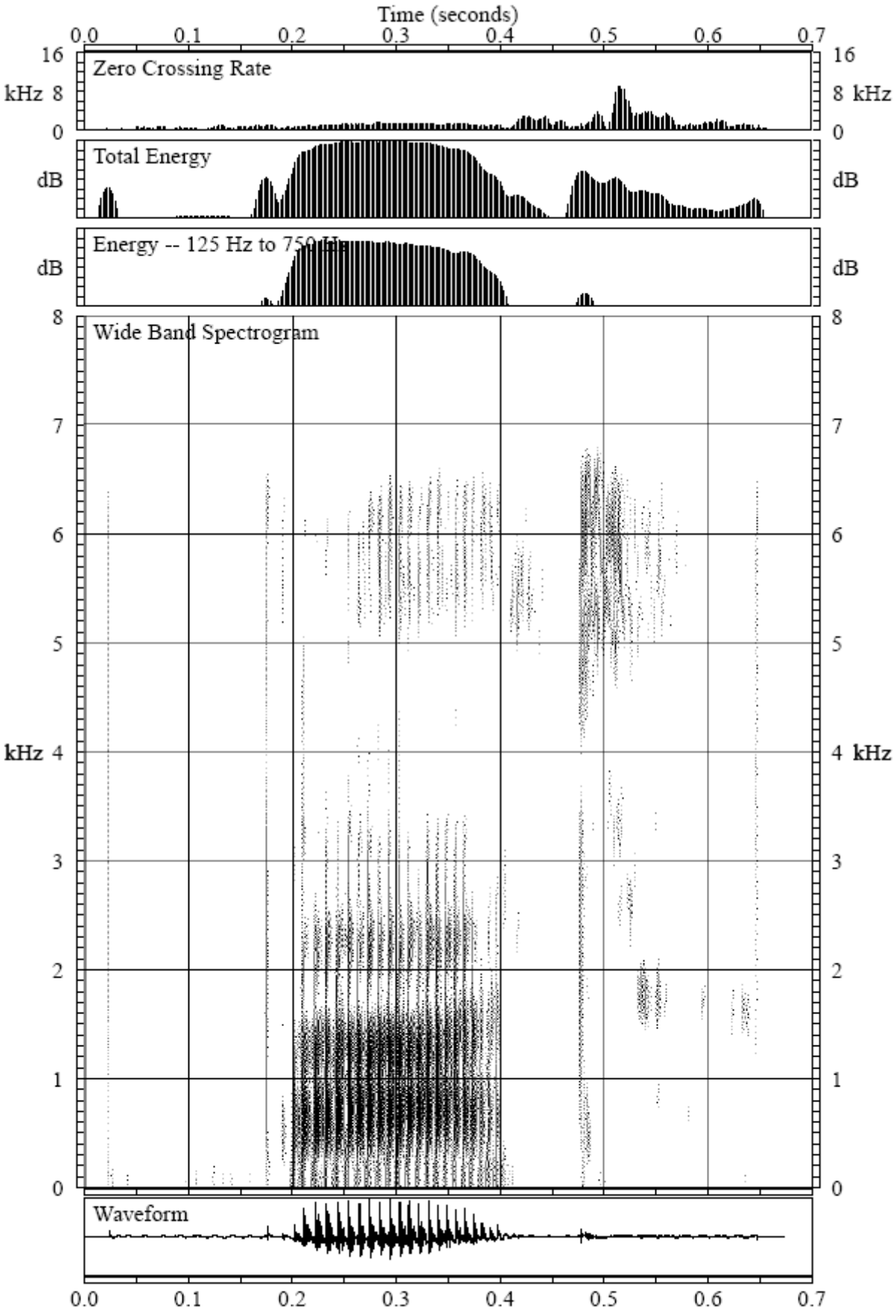


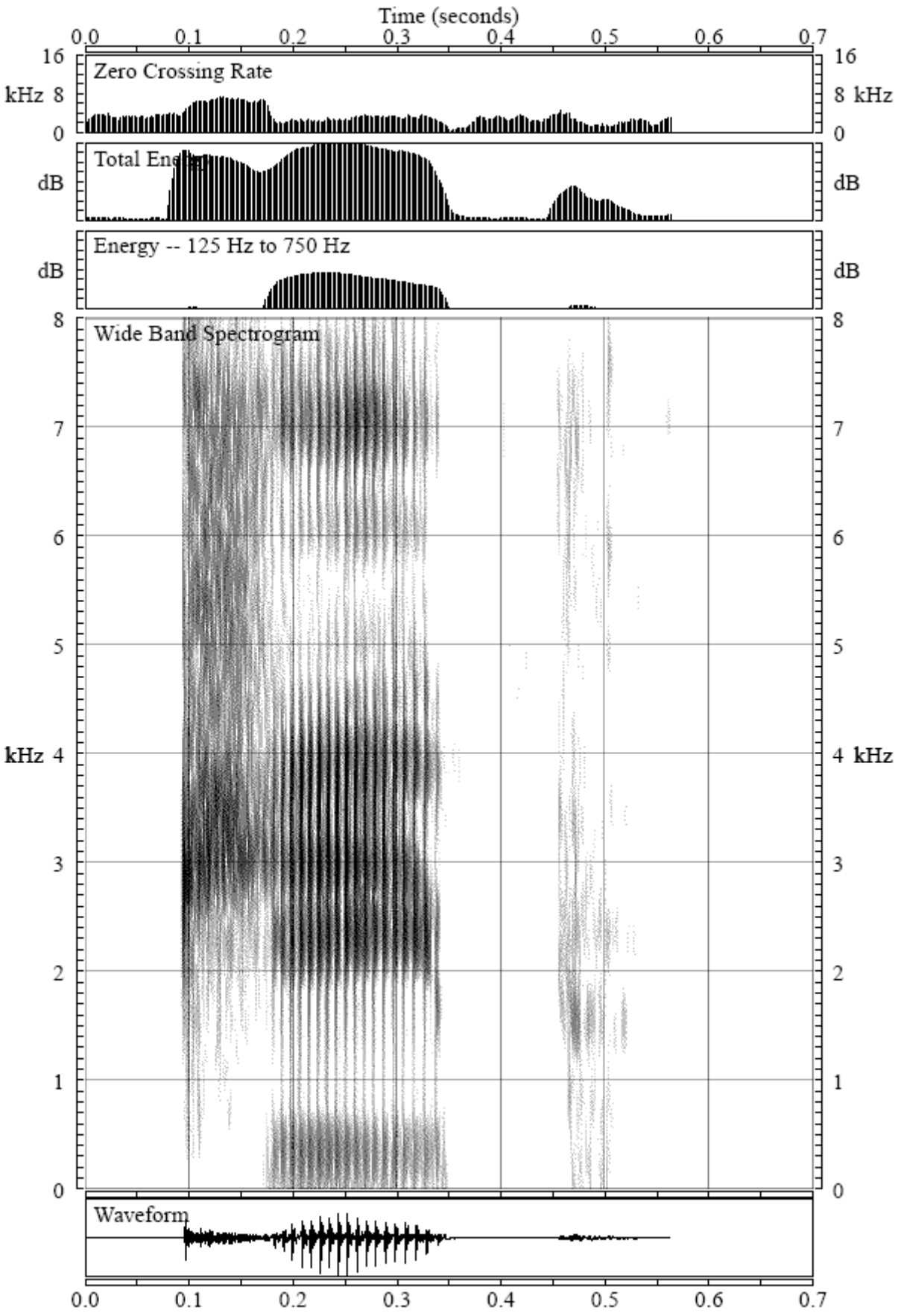
Рис. 1.7. Пример широко и узкополосного анализа Фурье речевого сигнала, F_s – частота квантования, T_w – длина окна анализа [4].


Time (seconds)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Zero Crossing Rate
16
kHz 8
0
Total Energy
dB
Energy -- 126 Hz to 750 Hz
dB
Wide Band Spectrogram
8
7
6
5
kHz 4
3
2
1
0
Waveform

/bit/


Time (seconds)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Zero Crossing Rate
16
kHz 8
0
Total Energy
dB
Energy -- 125 Hz to 750 Hz
Wide Band Spectrogram
8
7
6
5
kHz 4
3
2
1
0
Waveform

/bat/


Time (seconds)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Zero Crossing Rate
16
kHz 8
0
Total Energy
dB
Energy -- 125 Hz to 750 Hz
dB
Wide Band Spectrogram
8
7
6
5
kHz 4
3
2
1
0
Waveform

/kiʸp/

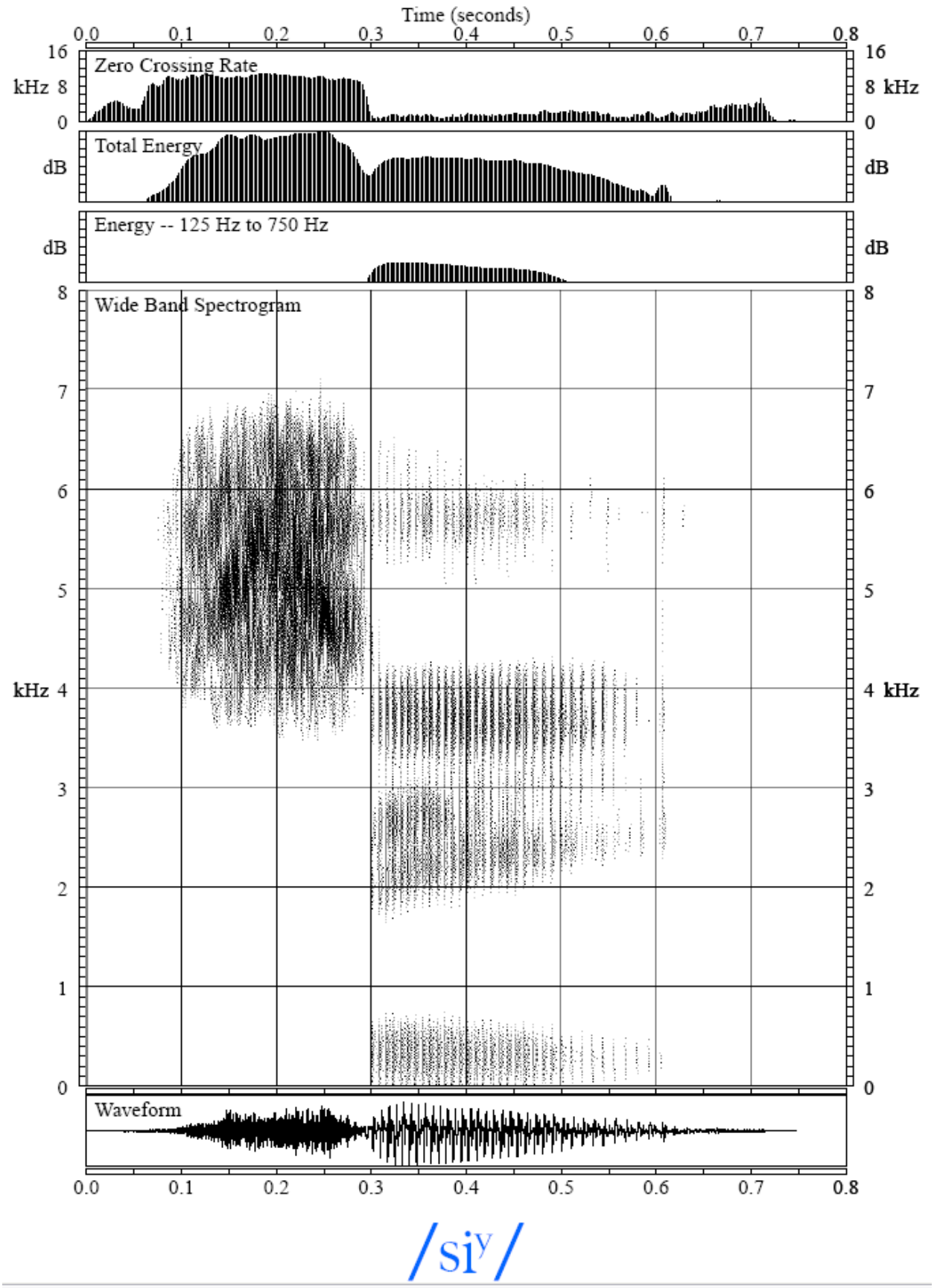


Рис. 1.8. Примеры спектрограмм и огибающих речевого сигнала [2].

В заключение данного раздела рассмотрим качественную сторону речеобразования и обсудим понятие «фонема».

Первоначально введенный Ф. де Соссюром в 1879 году термин «фонема» (phoneme) практически не отличался от языковедческого термина «звук», как единицы речи, подвергающейся научному анализу. Современное понимание этого термина фонетистами ближе к определению, данному Бодуэном де Куртенэ: «Фонема есть цельное, неделимое во времени представление звука языка».

Отметим два момента:

1. Фонема не есть физическая реализация звука, а является представлением звука в сознании (абстракцией).

2. Фонема воплощает идею атомарности, примененную к субъективному представлению о речи.

Явная неконструктивность данного определения с точки зрения технической реализации всегда вызывала споры и многочисленные варианты фонетической классификации. Технические специалисты, используя термин «фонема», вкладывали в него свои представления о речеобразовании и восприятии, доводя понятие до вульгарного: «фонема – это то, что я могу определить и выделить на своем приборе». Надо отметить, что и среди самих фонетистов нет единства в представлении о фонеме: московская и ленинградская (петербургская) школы фонетики предлагали алфавиты фонем, существенно отличающиеся по размеру. Наверное, для того, чтобы избежать этих конфликтов, технические специалисты ввели термин «фон» (phone) – конкретная реализация фонемы. Фоны, принадлежащие к одной фонеме, называются аллофонами.

Идея атомарности была очень привлекательной с технической точки зрения: достаточно установить характеристики или признаки составляющих речь «атомов» или «кирпичиков», и задача автоматического распознавания речи решена. Поиски признаков и выделение инвариантных к диктору и контексту «фонем» продолжались вплоть до 90-х, но успеха не имели (в том смысле, что ни одна из систем распознавания речи, насколько нам известно, результаты этих изысканий не использует).

Попробуем объяснить отсутствие успеха в поиске локализованных во времени фонем, рассмотрев процесс речеобразования на качественном уровне.

В процессе речеобразования речевые органы человека – губы, язык, нижняя челюсть, небная занавеска совершают движения, скорость которых зависит от силы мышц и массы самого органа. Скорость эта близка к предельной (иначе скороговорки не были бы популярны). Исследования показывают, что в естественной речи органы практически никогда не занимают положений, характерных для изолированно произнесенных звуков, а лишь обозначают движение в нужном направлении, соответственно, и форманты обозначают движение в нужном направлении. Очевидно, что движение в сторону, характерную для данной фонемы, зависит от предшествовавших и, как показывают эксперименты, даже последующих фонем, то есть, речевой аппарат может готовиться к произнесению некоторых звуков заранее. Этот эффект называется коартикуляцией. Взаимовлияние фонем не ограничивается соседями,

а может распространяться на несколько соседних фонем. Можно вычислить, насколько огромное число возможных сочетаний, скажем, хотя бы по три фонемы, существует для языка. Если к этому добавить нерешенную проблему признаков, инвариантных к диктору, то число возможных представлений фонемы становится настолько огромным, что о выуживании признаков вручную не может быть и речи.

Таким образом, по-прежнему используя аналогию с атомами, а лучше с квантами, можно заметить, что фонема скорее имеет «волновую» природу, то есть ее признаки «размазаны» по протяженному во времени отрезку, причем признаки различных фонем накладываются друг на друга. В разделе «Распознавание речи» будет рассказано, как эту проблему пытаются решить с помощью теории вероятностей и искусственных нейронных сетей.

Для того чтобы минимизировать вариативность фонов, вводят понятия “трифон” и “бифон”. Трифон, это фон, для которого определены предыдущая и последующая фонемы. Для бифона определена только одна из соседних фонем. Бифоны используют в начале и конце фрагментов речи, или когда данных для надёжной оценки трифона недостаточно. Легко подсчитать, что если количество используемых фонем равно N, то количество трифонов будет равно N^3. Аналогично вводят понятие “пентафон” для пяти последовательных фонем, однако для оценки параметров пентафонов требуются столь огромные базы данных, что в системах распознавания речи они практически не используются.

Все аспекты проблем, связанных с оценкой и использованием трифонов и бифонов будут рассмотрены в разделах 6, 7, 8, посвящённых использованию марковской модели для распознавания речи.

2. СЛУХОВАЯ СИСТЕМА

2.1. Строение уха человека

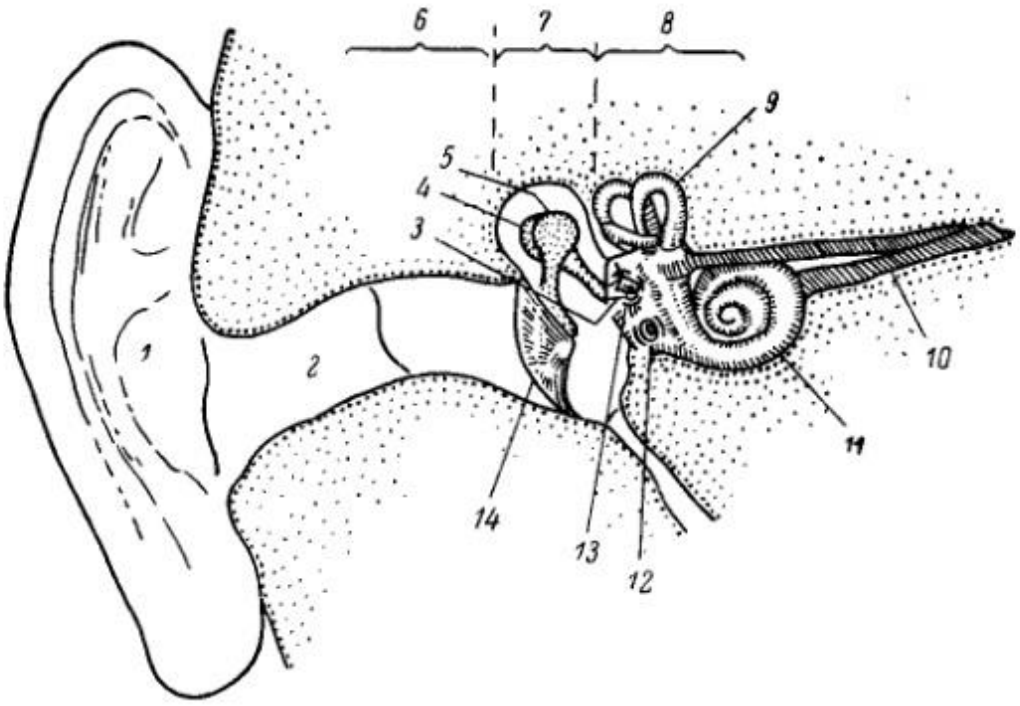
Изучение слухового анализатора важно для задачи автоматического распознавания речи, поскольку неадекватная обработка сигнала может привести к потере части полезных признаков и к излишне подробному представлению другой части.

В данном разделе ограничимся поверхностными сведениями о работе периферической слуховой системы, достаточными для обсуждения вопроса о том, какие параметры звукового сигнала способна выделять слуховая система.

Традиционно периферическую слуховую систему представляют как последовательно включенные наружное, среднее и внутреннее ухо (рис. 2.1). Наружное ухо – это звуковой проход к среднему, среднее – это система слуховых косточек, передающих звуковые колебания к внутреннему уху. Поскольку диаметры каналов невелики, звуковую волну в интересующем нас диапазоне частот можно рассматривать как плоскую. Амплитудно-частотные характеристики системы наружного уха приведены на рис. 2.2. Виден максимум в районе 3–4 кГц. Этот максимум обеспечивает наибольшую чувствительность слуховой системы в данном диапазоне.

Внутреннее ухо представляет собой очень интересный, с точки зрения спектрального анализа, прибор. Это конусная, трехполостная, разделенная мембраной (базилярной) вдоль трубка, заполненная перилимфой, жидкостью, близкой по свойствам к плазме крови. Длина улитки около 35 мм, и она, в соответствии с названием, закручена в спираль, делая 2.5 оборота. Как и в случае вокального тракта, поперечными колебаниями в расчетах пренебрегают и считают, что «закрученность» не оказывает влияния на распространение колебаний. Интересно, что подобную конструкцию, в целях экономии места, уже давно скопировали в радиолокационных системах.

Скорость распространения звука в улитке замедляется по сравнению с воздухом почти в 100 раз, так что задержка между возбуждением на входе в улитку и в апикальном конце от одного стимула составляет 7–8 мс. Базилярная мембрана улитки содержит около 3–4 тыс. «внутренних волосковых клеток», которые вырабатывают нервные импульсы в ответ на возбуждение. Возбуждение вызывается деформацией мембраны в процессе колебаний, но только в одном направлении. Таким образом, внутренние волосковые клетки осуществляют однополупериодное выпрямление сигнала. Можно сказать, что выпрямление «мягкое», поскольку во время второй половины колебания волосковые клетки уменьшают спонтанную активность. Описанный механизм действует до частот в 3–4 кГц. Далее на возбуждение волосковые клетки отвечают повышением общей активности. Как и нейроны, внутренние волосковые клетки имеют некоторый уровень насыщения, то есть не могут генерировать в единицу времени сколь угодно большое количество импульсов,



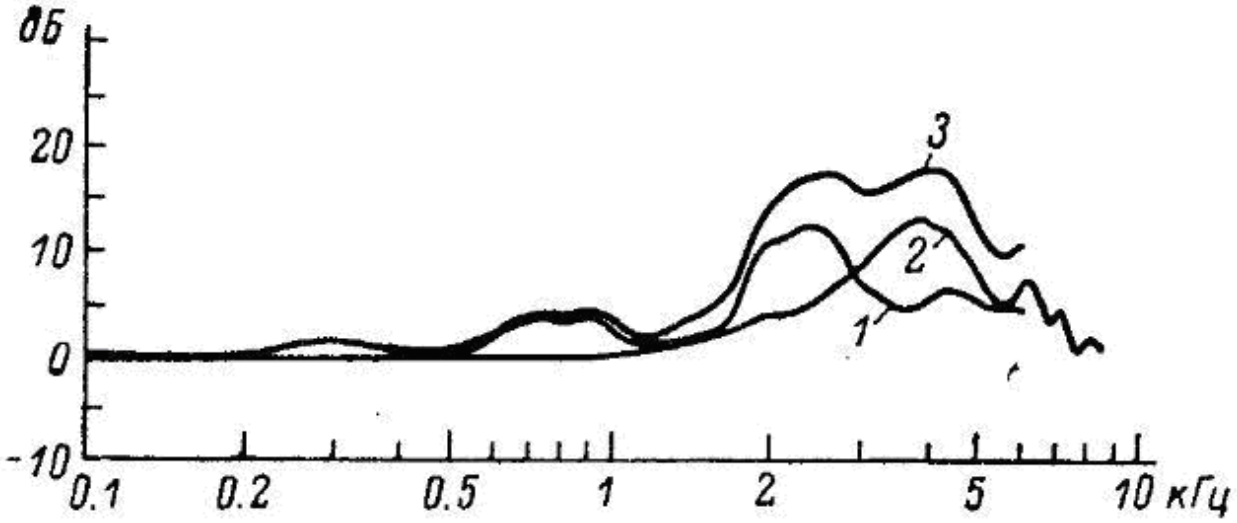
1 — ушная раковина; *2* — наружный слуховой проход; *3*, *4*, *5* — слуховые косточки, соответственно стремечко, наковаленка и молоточек; *6* — наружное ухо; *7* — среднее ухо; *8* — внутреннее ухо; *9* — вестибулярный аппарат; *10* — слуховой нерв; *11* — улитка; *12* — круглое окно; *13* — овальное окно; *14* — барабанная перепонка.

Рис. 2.1. Строение периферической слуховой системы [1].

и поэтому осуществляют компрессию сигнала. Существует предположение, подкрепленное экспериментальными данными [1], что улитка – это активная система, то есть, в ответ на звуковое возбуждение она начинает совершать колебания с энергией, большей, чем ей доставляется извне звуковой волной. Ответственность за этот механизм возлагают на «наружные волосковые клетки», которые также размещаются на базилярной мембране и которых в несколько раз больше, чем внутренних. Активность улитки проявляется только на малых уровнях возбуждения; видимо, это механизм нормализации уровня звука.

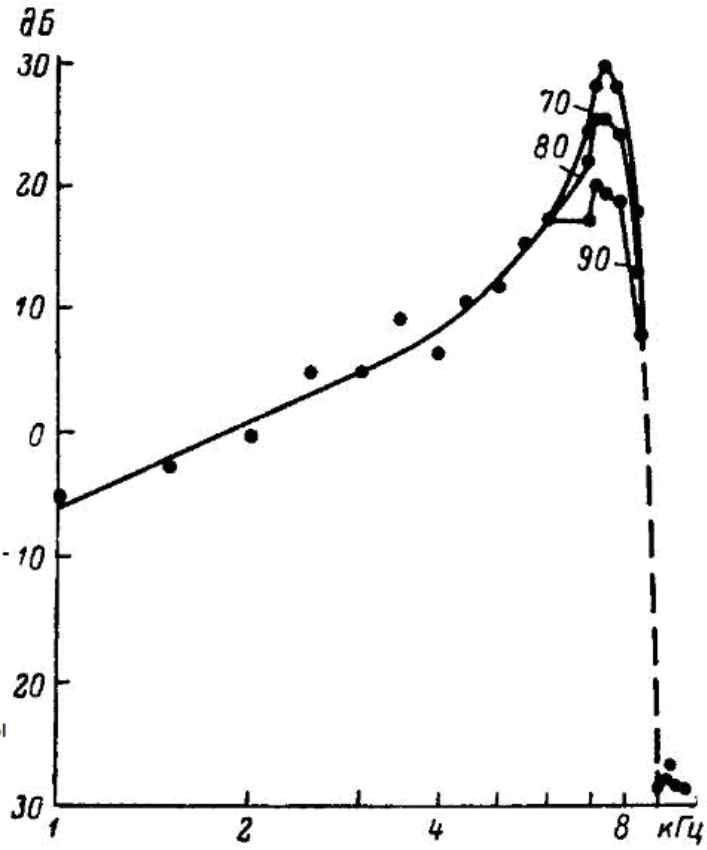
Не вдаваясь в хитросплетения восходящих нервных путей, поскольку это выходит за рамки рассматриваемой темы, отметим одну важную особенность слуховой системы – ее «тонотопическую» организацию. Этот термин означает, что области возбуждения, соответствующие близким частотам звукового стимула, находятся в близких (в прямом топологическом смысле) областях базилярной мембраны и сохраняют эту близость вплоть до высших разделов слуховой системы. Очевидно, данный факт является самым веским аргументом в пользу рассмотрения слуховой системы как спектрального анализатора.

Интересной особенностью улитки является то, что колебания различных частот проникают в нее на разную глубину. Высокочастотные (10 кГц и более) очень быстро затухают около входа в улитку. Чем ниже частота, тем дальше проходит сигнал. Сигналы с частотой 800 Гц и ниже проходят до конца улитки.



1 – амплитудно-частотная характеристика ушной раковины, определяемая как $P_0(w)/P_{вх}(w)$, 2 – наружного слухового прохода $P_б(w)/P_0(w)$, 3 – всей системы наружного уха, $P_б(w)/ P_{вх}(w)$.

Рис. 2.2. Амплитудно-частотные характеристики системы наружного уха [6]



По оси ординат - отношение амплитуды колебаний базилярной мембраны к амплитуде колебаний стремечка (в дБ). Цифры у кривых - уровни звукового давления у барабанной перепонки (в дБ).

Рис. 2.3. Зависимость амплитудно-частотной характеристики точки базилярной мембраны улитки мартышки от уровня звукового давления входного сигнала у барабанной перепонки [6]

Амплитудно-частотная характеристика элемента улитки с некоторой характеристической частотой имеет пологий подъем (6 дБ/окт) со стороны низких частот (то есть все частоты ниже характеристической представлены в колебании данного участка базилярной мембраны, но с уменьшающимся весом) и резкий спад (по различным данным от 100 до 300 дБ/окт) в сторону высоких частот (рис. 2.3.). Таким образом, главный элемент периферической слуховой системы, не обладая сильными резонансными свойствами, представляет скорей линию задержки или временной анализатор.

2.2. Маскировка. Восприятие высоты звука

Под маскировкой понимают повышение порога слышимости звука (стимула) в присутствии других звуков (маскеров). Маскирующие звуки могут предшествовать (прямая), действовать одновременно (одновременная) и следовать за сигналом (обратная маскировка).

Стимулы и маскеры могут иметь различную структуру. Обычно используют узкополосный шум, либо тональные посылки. Результаты экспериментов довольно трудно распространить на произвольные сигналы, поскольку для процессов слуховой обработки характерны нелинейности. Эксперименты с тональными стимулами и маскерами позволили построить кривые, определяющие пороги слышимости по частоте и времени, которые использовались при разработке алгоритмов сжатия MPEG.

Рассмотрим два эксперимента по маскировке, которые дают представление об особенностях восприятия звука на разных частотах.

Зафиксируем частоту стимула и будем маскировать его узкополосным шумом с центром на частоте стимула, постепенно расширяя полосу маскера, оставляя неизменной спектральную плотность энергии. Оказывается, что, начиная с некоторой ширины, дальнейшее расширение маскера по спектру не приводит к существенному увеличению порога восприятия, то есть незначительно усиливает маскировку. Ширина такой полосы как функция частоты называется критической полосой. Результат можно грубо интерпретировать так, как если бы в слуховой системе присутствовали близкие к прямоугольным или трапециевидным частотные фильтры, в границах которых стимулы суммируются, при этом стимулы, разделённые более чем на критическую полосу, обрабатываются независимо. В диапазоне от 100 Гц до 16 кГц насчитывается 24 критические полосы (см. табл. 2.1.). Шкала критических полос называется шкалой «барк».

Второй эксперимент состоит в том, что слушателю предъявляют гармонический сигнал и просят выставить частоту второго сигнала так, чтобы на его субъективный взгляд она была в два раза выше или ниже, чем частота предъявленного. Полученная таким образом шкала называется шкалой «мел» (см. табл. 2.1.).

Таблица 2.1.

Шкалы частот барк и мел [4]

Index	Bark Scale		Mel Scale	
	Center Freq. (Hz)	BW (Hz)	Center Freq. (Hz)	BW (Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969

Частоты в барках и мелах можно рассчитывать по следующим формулам:

$$Bark(f) = 13\arctan(0.00076f) + 3.5\arctan\left((\frac{f}{7500})^2\right).$$
$$Mel(f) = 1125\ln(1+\frac{f}{700}). \quad (2.1)$$

Восприятие высоты звука зависит от его интенсивности (рис. 2.4.), поэтому, для получения однозначных результатов, интенсивность стимулов во втором эксперименте фиксируют в 40 дБ.

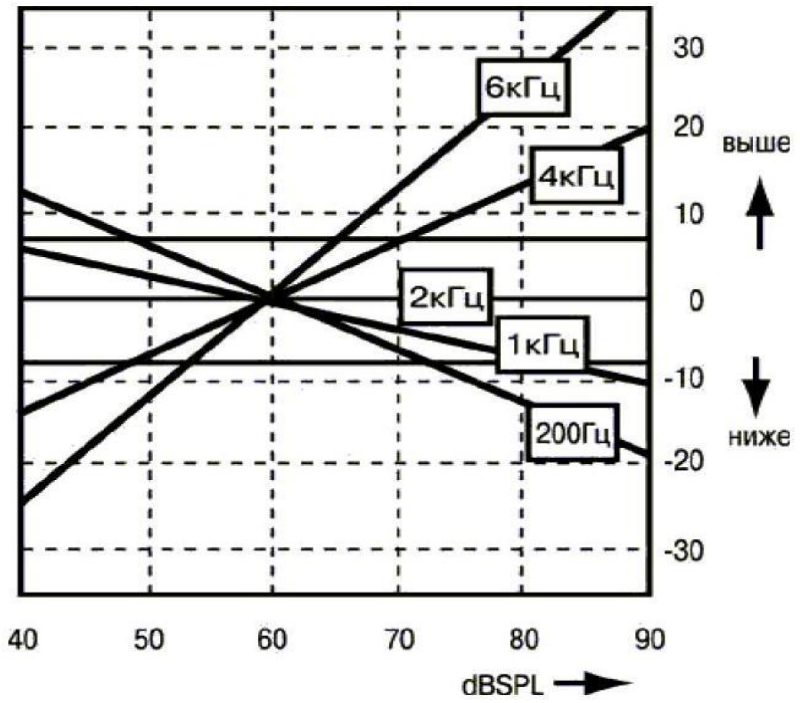


Рис. 2.4. Зависимость высоты звука от его интенсивности [5] .

Из данных таблицы 2.1 следует, что слуховая система склонна рассматривать низкочастотные компоненты речи более подробно, чем высокочастотные – начиная с 1000 Гц, шкалы можно считать близкими к логарифмическим. Отсюда следует, что при обработке речи для распознавания можно сэкономить на представлении высоких частот в наборе признаков.

2.3. Восприятие громкости звука. Кривая равной громкости

Чувствительность слуховой системы к различным частотам различна. Из всего диапазона 20–20000 Гц самые низкие пороги восприятия относятся к диапазону 2–5 кГц, в основном, благодаря передаточной функции слухового канала (рис. 2.2.).

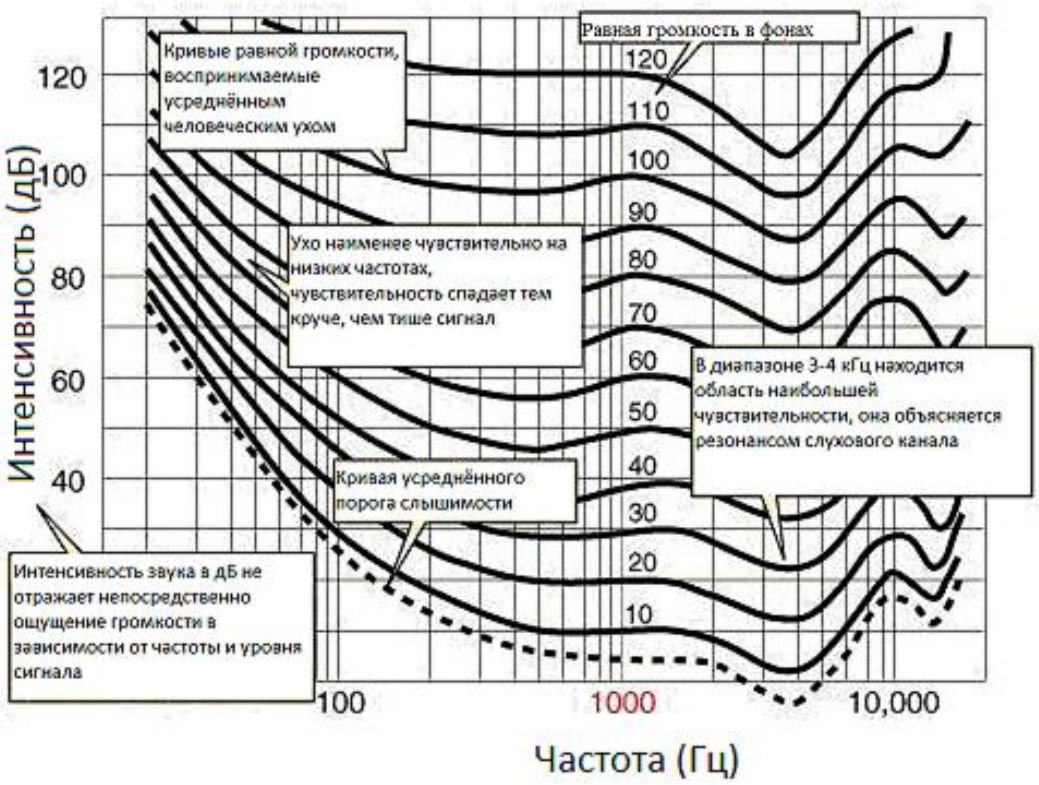


Рис. 2.5. Кривые равной громкости (http://hyperphysics.phy-astr.gsu.edu/hbase/sound/eqloud.html).

Громкость – это субъективная оценка интенсивности звука. Ощущение громкости зависит не только от частоты, но и от длительности звукового стимула. Для количественной оценки абсолютной громкости была принята специальная единица сон. Громкость в 1 сон – это громкость синусоидального звука с частотой 1000 Гц и уровнем 40 дБ относительно звукового давления $2 \cdot 10^{-5}$ Па. Количественно зависимость воспринимаемой громкости звука ***S*** (в сонах) и его звукового давления может быть представлена в следующем виде:

$S = Cp^{0.6}$, где C – постоянная, зависящая от частоты сигнала.

Отсюда следует, что при увеличении звукового давления на 10 дБ, громкость возрастает в 2 раза. Таким образом, зависимость между звуковым давлением и ощущением громкости носит логарифмический характер.

2.4. Адаптация

Реакция слуховой системы на продолжительный стимул начинается резким всплеском, далее следует спад с выходом на постоянный уровень. При выключении стимула система на некоторое время снижает спонтанный уровень импульсации (рис. 2.6.).

На рис. 2.7 представлена простейшая модель адаптации. Здесь блок 1 – безынерционная компрессионная нелинейность, например:

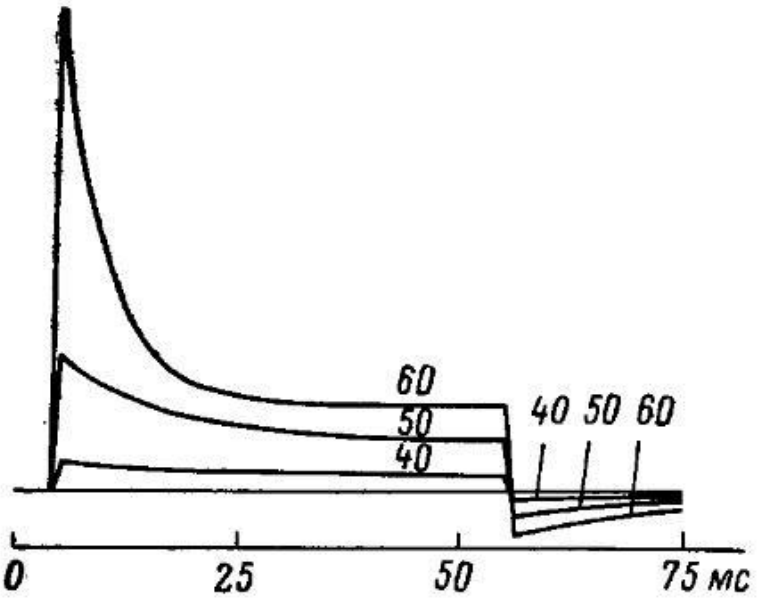


Рис. 2.6. Огибающая реакции на посылку тона с прямоугольной огибающей. Параметр кривых – интенсивность входного сигнала в дБ от условного порога [6].

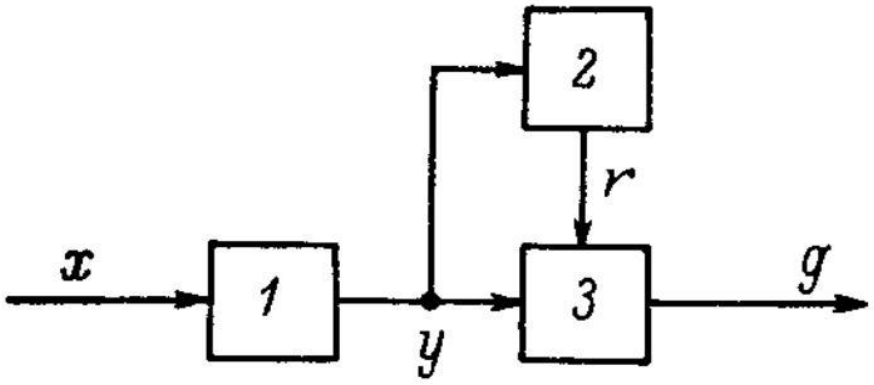


Рис. 2.7. Простейшая модель адаптации [6].

$y = (1 + A\log\frac{x}{x_0})^p$ при $x>=x_0$, A и p – константы, использовались 0,5 и 4,6,

$y = x$ при $0<=x<=x_0$,

$y = 0$ при $x<0$;

блок 2 – фильтр нижних частот (интегратор):

$\tau\frac{\partial r}{\partial t} + r = y$, или в дискретном виде: $r_i = (1-\frac{T}{\tau})r_{i-1} + \frac{T}{\tau}y_i$, где T – период квантования, τ – постоянная времени;

блок 3 – блок переменного коэффициента передачи:

$g = \frac{y}{\alpha + \beta r}$, где α и β – константы.

Адаптация является важным свойством всех физиологических систем. При восприятии звуков адаптация позволяет повысить помехоустойчивость сообщения за счёт подавления стационарных шумов.

2.5. Физиологические методы обработки сигналов

Возникает вопрос: каким образом, обладая столь низкой спектральной избирательностью на уровне каждого участка базилярной мембраны, слуховая система в целом демонстрирует очень высокую избирательность? Ведь эксперименты по восприятию тональных стимулов, а также характер музыки, сформировавшейся у всех народов, наличие людей с абсолютным слухом и присутствие хоть какого-то слуха у остальных не оставляют сомнения, что разрешающая способность по спектру слуховой системы достаточно высока. Для объяснения этого феномена привлекается уже упомянутый факт активного возбуждения улитки. Кроме того, обострение спектра можно объяснить, используя нелинейные методы параллельной обработки, присущие нервной системе вообще. Исключительно для того, чтобы стало понятно, какие интересные возможности предоставляют эти методы, приведём следующий механизм:

В 1984 г. С. Сенеф [7] предложила, как тогда казалось, довольно искусственный, но эффективный метод обострения спектральных максимумов для гребенки фильтров (рис. 2.8.). Метод заключался в том, что в каждом спектральном канале после однополупериодного детектирования делалось ответвление, сигнал в котором задерживался на половину периода $1/(2*Fc)$, где Fc – характеристическая частота данного канала. После этого сигнал в задержанном канале вычитался из сигнала в основном канале с последующим детектированием и сглаживанием фильтром нижних частот (ФНЧ). Очевидно, что если частота сигнала равна характеристической частоте данного фильтра, то ослабления сигнала не произойдет, поскольку сдвинутый на половину периода сигнал будет вычитаться в те отрезки времени, когда в основном канале сигнал и так равен нулю по причине предварительного детектирования. В остальных каналах сигнал будет ослабляться. Понятно, что этот метод позволяет добиться более высокой разрешающей способности гребенки фильтров без увеличения времени переходных процессов. Покажем, что этот метод может реализоваться в периферической слуховой системе без привлечения искусственного построения, связанного с ответвлением и задержкой в каждом канале. Действительно, в соответствии с рассмотренным механизмом работы улитки, каждая частотная компонента присутствует во всех областях улитки и в

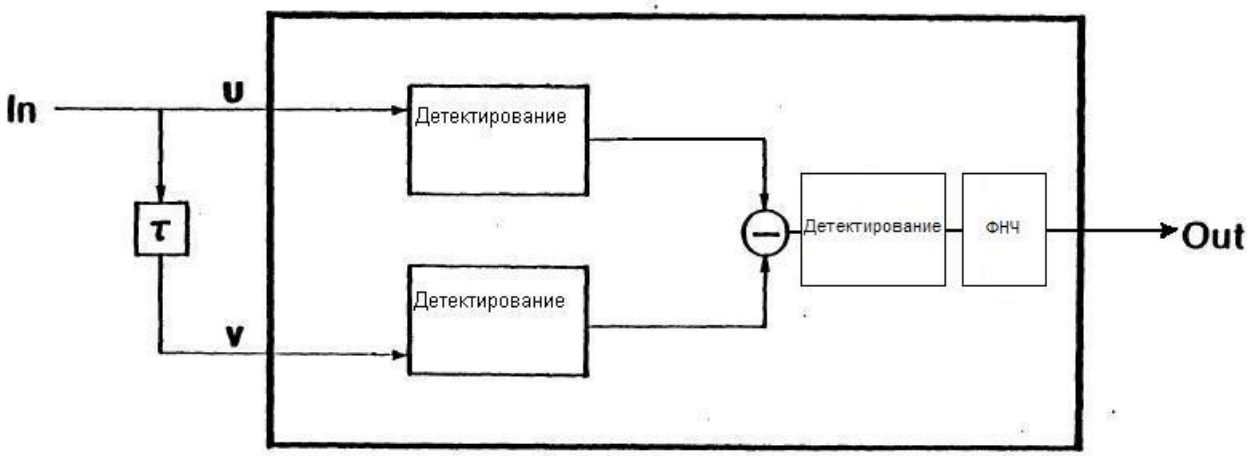


Рис. 2.8. Обобщённый синхронный детектор [7], τ = 1/(2Fc).

соответствующих нервных каналах для участков базилярной мембраны с характеристическими частотами выше рассматриваемой, то есть, ближе к входу в улитку (рис. 2.9.).

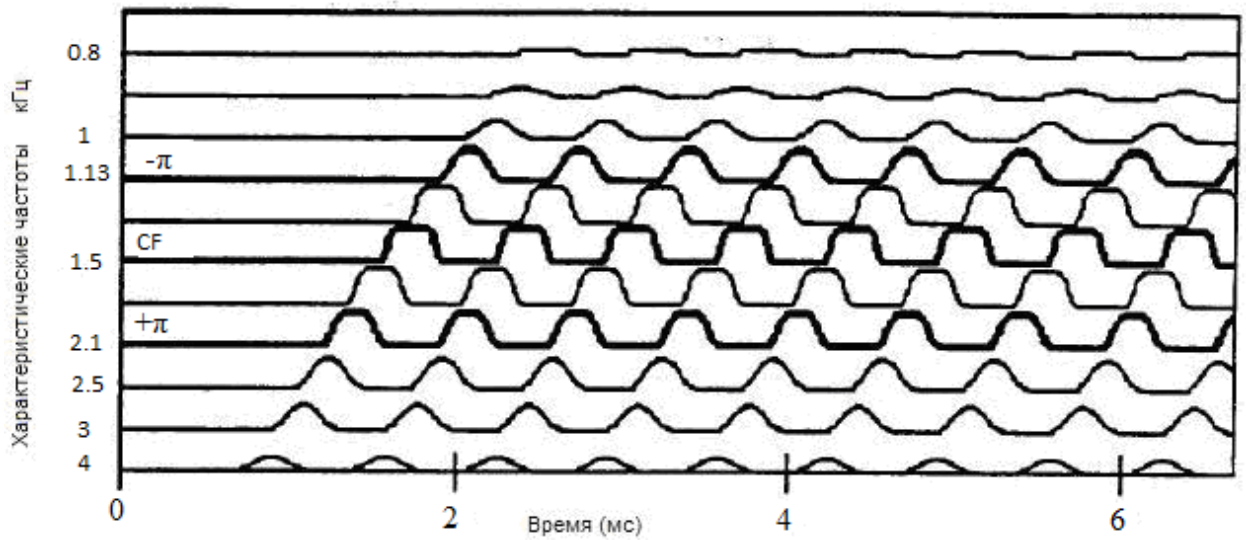


Рис. 2.9. Пространственно-временной отклик слуховой системы на чистый тон 1,5 кГц

Вопрос в том, как подводится сигнал, задержанный на половину периода относительно характеристической частоты к данному каналу? В нервной системе существует универсальный механизм «латерального торможения», согласно которому сигнал в каждом канале собирает возбуждения в непосредственной близости от себя со знаком плюс и вычитает сигналы, проходящие по каналам в некотором отдалении. Весовую функцию, как функцию от частоты, или, в соответствии с принципом тонотопической организации, расстояния от данного канала, можно представить как разность двух Гауссовых функций с различной дисперсией (рис. 2.10). Таким образом, каждый канал своим сигналом старается подавить соседей, которые соответствуют минимуму весовой функции латерального торможения, и,

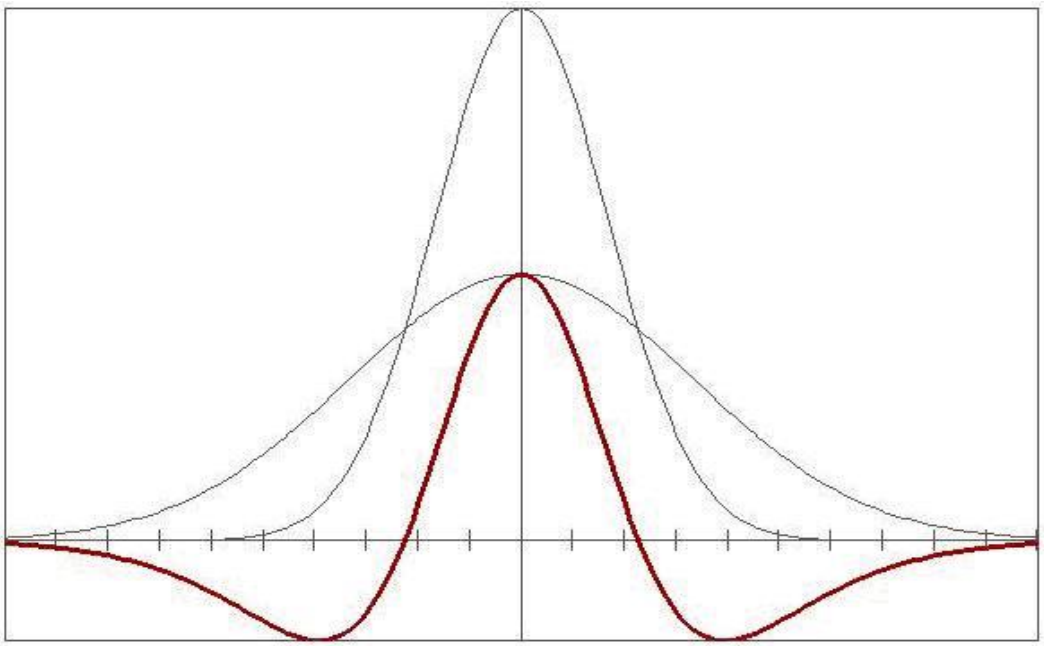


Рис.2.10. Весовая функция латерального торможения

аналогично, подавляется какими-то другими каналами. Остается убедиться, что для улитки минимумы весовой функции латерального торможения могут соответствовать сигналу, сдвинутому на половину периода характеристической частоты. Заметим, что низкочастотная ветвь весовой функции роли не играет, поскольку она соответствует участкам базилярной мембраны, удаленным от входа по отношению к рассматриваемому участку, а туда сигнал с рассматриваемой характеристической частотой не доходит по причине быстрого затухания (рис. 2.3). Экспериментальные данные показывают, что характерные для улитки ширины спектральных полос каналов полностью соответствуют описанному механизму, то есть, минимум весовой функции латерального торможения для участка базилярной мембраны с рассматриваемой характеристической частотой вполне может приходиться на участок мембраны, для которой время прихода сигнала меньше на половину периода данной характеристической частоты. Механизм компрессии или ограничения амплитуды усиливает обостряющий эффект данного механизма, поскольку основной и задержанный сигналы сближаются по амплитуде, и подавление сигналов в каналах происходит более эффективно.

Попытаемся понять, каким образом в нейронных сетях слуховой системы могут возникать довольно экзотические весовые функции типа латерального торможения. Надо отметить, что в зрительном восприятии были обнаружены весовые функций ещё более изощрённого вида, названные «габоровскими» по имени их первооткрывателя. Габоровские функции напоминают двумерные вейвлеты.

Основой нейронных сетей является нейрон, упрощённая модель которого изображена на рис.2.11 В качестве функции активации F(s) могут использоваться:

1. линейная функция,
2. пороговая функция,

3. сигмоидальная функция (1/(1+exp(-(ks-a)), гиперболический тангенс,…) рис.2.12.

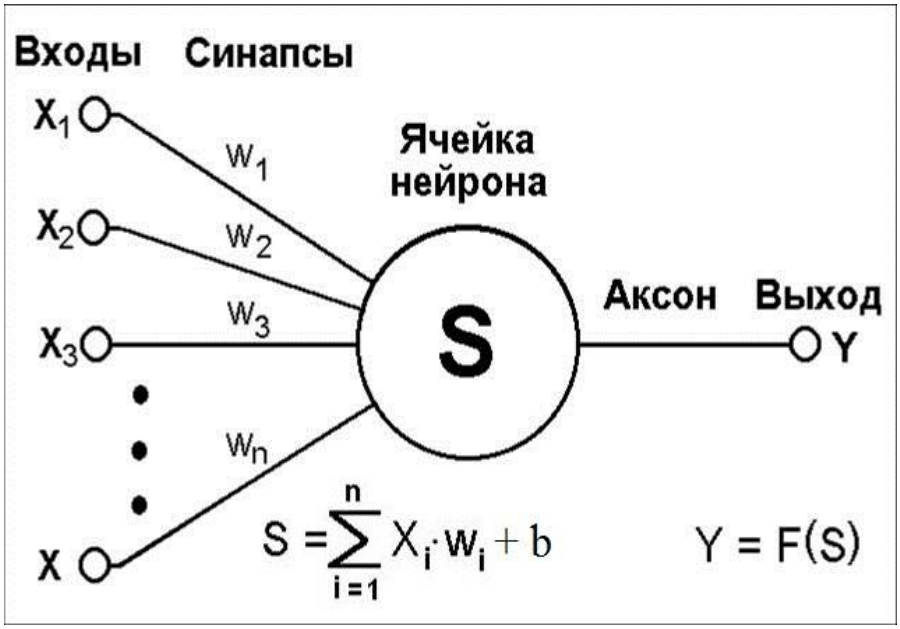


Рис. 2.11. Модель нейрона

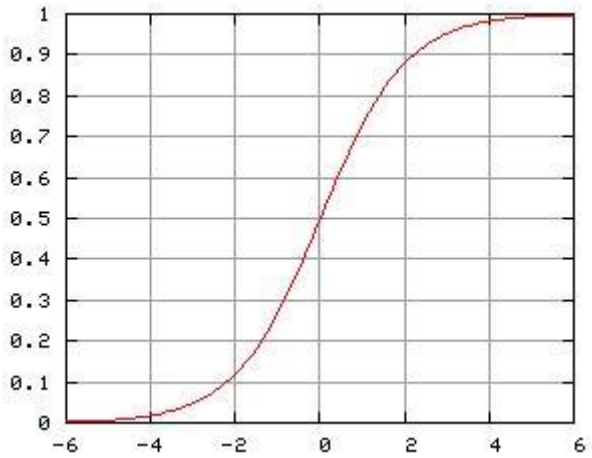


Рис. 2.12. Сигмоидальная функция

Нейроны могут объединяться в нейронные сети самым произвольным образом. Наиболее распространённой структурой в моделях является «перцептрон» – несколько плоских слоёв нейронов, в которых каждый нейрон собирает сигналы со всех нейронов предыдущего слоя и передаёт на все нейроны следующего. Связи между нейронами одного слоя и обратные связи запрещены. Подробнее о нейронных сетях будет рассказано в разделах, посвящённых современным методам распознавания. В слуховой системе нейроны первого слоя собирают информацию с волосковых клеток на некотором протяжении базилярной мембраны. Плотность синапсов (см. рис. 2.11), то есть их количество на единицу длины базилярной мембраны, убывает по мере удаления от нейрона приблизительно в соответствии с гауссовой функцией. Ширина захвата нейронов,

или дисперсия этой гауссовой функции, различна для различных нейронов. Допустим, что все входы некоторых нейронов первого слоя возбуждающие. Уже на втором слое (первом скрытом) могут быть нейроны, которые суммируют сигналы от нейронов первого слоя с различными знаками. Если ширины захвата этих нейронов различны, то произойдёт вычитание гауссовых функций, что и приведёт к образованию весовой функции типа латерального торможения относительно сигналов волосковых клеток. Нетрудно представить, что на следующих слоях нейронной сети с помощью этого же механизма можно получить весовые функции типа габоровских или вейвлетов. Также очевидно, что нейроны всё более высоких слоёв получают информацию о колебаниях всё более широких участков базилярной мембраны, то есть могут анализировать и сравнивать всё более широкие полосы спектра.

Данное рассмотрение не претендует на точность с точки зрения физиологии, но даёт представление о возможных методах обработки сигналов слуховой системой.

2.6. Выводы

Требования к признакам и процедурам обработки на основе первых двух глав руководства можно сформулировать следующим образом:

- использовать спектральный анализ с логарифмической шкалой;
- использовать логарифмическую компрессию громкости;
- разработать процедуру адаптации/АРУ;
- использовать совместно спектральные и временны**ы**е признаки;
- использовать совместно долговременный анализ (до 500 мс) и кратковременный анализ (20–25 мс).

Вернёмся к вопросу о признаках, которые содержат информацию, достаточную для автоматического распознавания речи. Хотя само изложение материала подсказывает, что этот вопрос уже решён в пользу спектрального преобразования (или связанных с ним методов), остановимся на нём подробнее.

То, что преобразование Фурье, как и любое другое обратимое преобразование, сохраняет информацию, заключённую в исходном сигнале, конечно, не является аргументом в пользу спектра Фурье. Важно, чтобы выделяемые признаки несли в каком-то смысле инвариантную к дикторам и внешним условиям информацию о словесном содержании сообщения и отбрасывали нерелевантную информацию. Отметим, что для распознавания используется только амплитудная составляющая спектра, то есть исходное количество информации безболезненно уменьшается в два раза, поскольку фазовая составляющая не воспринимается слуховой системой. К сожалению, используемые при распознавании речи признаки не реализуют в полной мере идею инвариантности и, по-видимому, никогда не смогут её реализовать, поскольку расщепление информации на словесное содержание, индивидуальность диктора и внешние шумы – это работа «высших уровней». Спектр Фурье, по-видимому, является единственным кандидатом на роль первичных признаков системы распознавания. Этот вывод представляется

естественным, если сопоставить параметры речевой системы, которые контролирует человек и параметры акустического сигнала, к которым слуховая система относится особенно «бережно». Вспомним, как происходит постановка произношения при обучении иностранным языкам – преподаватель обращает внимание на форму и вытянутость губ, положение нижней челюсти (зубов) относительно кончика языка, положение тела языка и положение нёбной занавески при произнесении назализованных звуков. Все эти факторы определяют форму голосового тракта, положение смычек и сужений, то есть параметров, определяющих текущую огибающую спектра сигнала. Слуховая система, в свою очередь, с помощью улитки производит спектральный анализ сигнала. Что особенно важно, информация о спектральных компонентах проходит до соответствующих отделов центральной нервной системы, не перемешиваясь и подвергаясь лишь некоторой очистке и обострению. Эта особенность периферической слуховой системы, получившая название «тонотопической организации», доказывает, что амплитудный спектр сигнала является основой для распознавания речи человеком и, следовательно, для автоматических систем распознавания речи. Аргументы, указывающие, что можно получить признаки абстрактными методами, не опираясь на изучение слуховой системы, поскольку человек научился выполнять некоторые функции живых существ лучше них, используя другие методы, например: «самолёты ведь не машут крыльями!», в данном случае, по-видимому, не проходят [8]. Надо учесть, что до сих пор человек моделировал взаимодействие живых существ с внешним миром в соответствии с хорошо изученными физическими законами. Очевидно, что человек полетел бы, даже если бы на Земле не водились птицы и прочие летающие, к которым он испытывал зависть. Аналогичное утверждение относительно распознавания речи полностью лишено смысла. Автоматическое распознавание речи является уникальной задачей моделирования системы, развившейся в процессе филогенеза за несколько сотен тысяч лет. В этой системе «передатчик» и «приёмник» сигналов управляются одним органом – мозгом, и в течение этих тысячелетий они нашли «общий язык», который и надо расшифровать. Очень сомнительно, что расшифровка допускает альтернативные варианты, тем более сомнительно, что они лучше «натуральных». Очевидным следствием этих рассуждений является также то, что перспективные системы распознавания речи должны в максимальной степени использовать достижения физиологии в области слухового анализа. Однако следует иметь в виду, что слепое копирование открытых механизмов восприятия может даже ухудшить распознавание, поскольку в живых системах механизмы обработки редко функционируют изолированно друг от друга. Скорее речь может идти об общих принципах обработки информации в живых системах – многоэтапной иерархической обработке с использованием большого количества нейронов.

3. ПРИЗНАКИ РЕЧЕВОГО СИГНАЛА ДЛЯ РАСПОЗНАВАНИЯ РЕЧИ

В данном разделе будут рассмотрены некоторые методы цифровой обработки речевых сигналов, в частности:

1. Спектр Фурье.
2. Спектр Фурье в шкале мел.
3. Выход гребёнки цифровых фильтров.
4. Коэффициенты линейного предсказания.
5. Кепстр.

Для получения спектра Фурье используется алгоритм БПФ с длиной окна, соответствующей 2–4 периодам основного тона, то есть около 20 мс. При частоте квантования 10–16 кГц окно обычно выбирается размером 256 отсчётов. Обычно окна сдвигают на 10 мс, обеспечивая частоту следования векторов признаков 100 Гц. Полезно использовать окно Хэмминга для ослабления искажений сигнала, вызванных применением к непрерывному сигналу конечного окна анализа, по формуле:

$$S'(n) = [0.54 – 0.46 * cos(2*PI*n / (N–1))]*S(n), \quad (3.1)$$

где $n = 1,\ldots,N$, N – размерность окна, $S(n)$ – отсчеты речевого сигнала.

Для перехода в шкалу мел спектральные отсчёты в границах каждой спектральной полосы суммируют с некоторыми коэффициентами, представляющими окно, обычно, треугольное (рис. 3.1.).

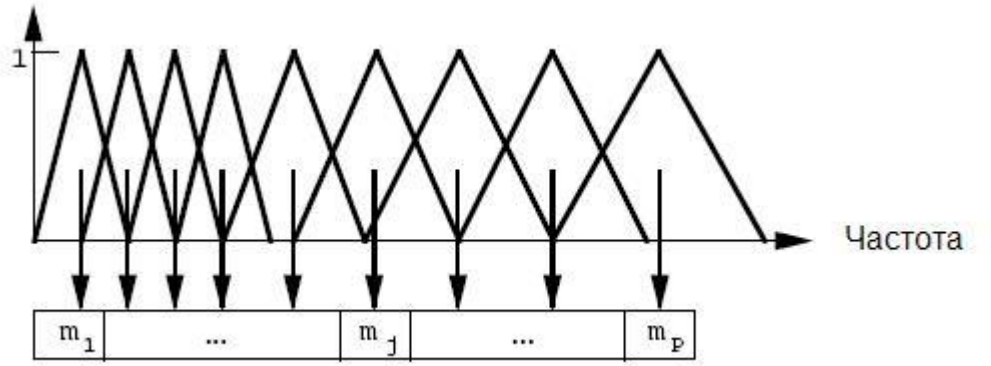


Рис. 3.1. Гребёнка фильтров для суммирования спектральных отсчётов

В качестве гребёнки цифровых фильтров используются рекурсивные фильтры второго порядка:

$$y_i = x_i + c_1 * y_{i-1} + c_2 * y_{i-2}, \quad (3.2)$$

где $c_1 = 2*R*cos(2\pi*fp/F)$, $c_2 = -R*R$, x_i – отсчёты входного сигнала, y_i – отсчёты выходного сигнала, F – частота квантования, fp – частота полюса фильтра, R – радиус полюса.

Ширина фильтра на уровне 0.5 от максимума по энергии приблизительно равна:

$$\Delta f \approx \frac{F}{2\pi}\frac{1-R}{\sqrt{R}}, \tag{3.3}$$

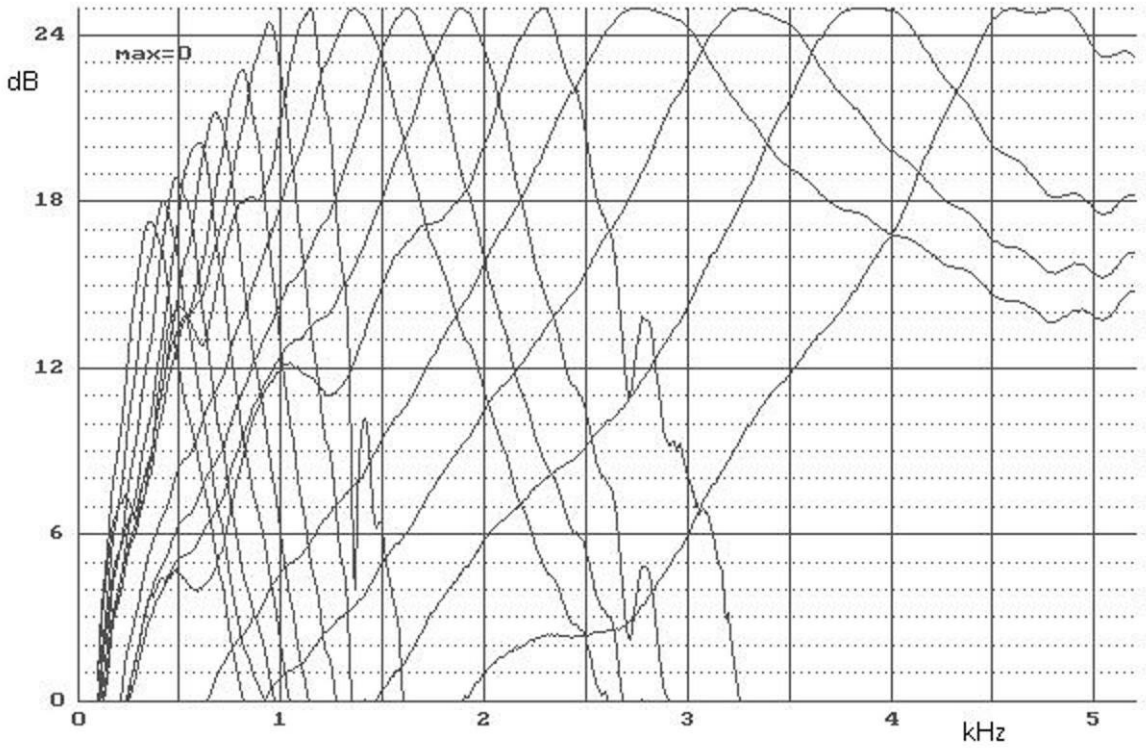


Рис. 3.2. Амплитудные характеристики гребёнки цифровых фильтров

для δ=1-R<<1, с точностью до величин первого порядка по δ:

$$\Delta f \approx \frac{F}{2\pi}(1-R). \tag{3.4}$$

Радиус полюсов фильтра R может равняться 1, что соответствует границе неустойчивости фильтра. В этом случае анализ аналогичен оконному: через фильтр пропускается необходимое количество отсчётов, оценкой спектра является сумма модулей или квадратов выходов фильтра y_i, после чего значения y обнуляются для приёма следующего фрагмента сигнала. Если радиус меньше 1, то с его помощью можно регулировать добротность, или ширину фильтра. Преимуществом гребёнки цифровых фильтров является то, что её можно сразу организовать в шкале мел. На рис. 3.2. изображены характеристики гребёнки цифровых фильтров, использовавшиеся в системе распознавания команд, разработанной в Центре Речевых Технологий.

Другим направлением спектрального анализа речи, отличным от Фурье анализа, является использование метода линейного предсказания. Уже в конце 70-х годов многие зарубежные системы распознавания строились на

использовании стандартных коэффициентов линейного предсказания (LPC). Модель линейного предсказания речи предполагает, что передаточная функция голосового тракта представляется полюсным фильтром с передаточной функцией [9]:

$$H(z) = \frac{1}{\sum_{i=0}^{p} a_i z^{-i}} \quad . \tag{3.5}$$

где p – число полюсов и $a_0 = 1$;

Фильтр с такой передаточной функцией позволяет описать поведение сглаженного спектра речевого сигнала с хорошей точностью, за исключением назализованных звуков. Коэффициенты фильтра $\{a_i\}$ – выбираются путем минимизации среднеквадратичной ошибки предсказания, просуммированной на окне анализа.

Автокорреляционный метод, позволяющий выполнить эту оптимизацию, заключается в следующем. Для заданного окна отсчетов речевых сигналов $\{s_n, n = 1,N\}$, первые $p+1$ членов автокорреляционной последовательности вычисляются по формуле:

$$r_i = \sum_{j=1}^{N-i} s_j s_{j+i} \, , \tag{3.6}$$

где $i = 0,\ldots,p$;

Коэффициенты фильтра вычисляют рекурсивно с использованием множества вспомогательных коэффициентов $\{k_j\}$, которые могут быть интерпретированы как коэффициенты отражения эквивалентной акустической трубы и ошибки предсказания E, которая изначально равна r_0. Пусть $\{k_j^{(i-1)}\}$ и $\{a_j^{(i-1)}\}$ будут коэффициентами отражения и коэффициентами для фильтра i-1 порядка соответственно, тогда фильтр порядка i может быть вычислен за три шага. Вначале вычисляется новое множество коэффициентов отражения:

$$k_j^{(i)} = k_j^{(i-1)} \tag{3.7}$$

для j от 1 до i-1 и

$$k_i^{(i)} = \left\{ r_i + \sum_{j=1}^{i-1} a_j^{(i-1)} r_{i-j} \right\} / E^{(i-1)} . \tag{3.8}$$

Затем пересчитывается энергия предсказания:

$$E^{(i)} = \left(1 - k_i^{(i)} k_i^{(i)}\right) E^{(i-1)} . \tag{3.9}$$

В заключение вычисляются новые коэффициенты фильтра:

$$a_j^{(i)} = a_j^{(i-1)} - k_i^{i} a_{i-j}^{(i-1)} \tag{3.10}$$

для $j = 1$, i-1 и

$$a_i^{(i)} = -k_i^{(i)}. \quad (3.11)$$

Этот процесс продолжается от i=1 до $i = p$, где p – требуемый порядок фильтра. Обычно используется p = 12.

Альтернативой параметрам, основанным на коэффициентах линейного предсказания, является кепстр, полученный из спектра Фурье или коэффициентов линейного предсказания (LPCC).

Кепстр линейного предсказания может быть эффективно вычислен с помощью простой рекурсии:

$$c_n = -a_n - \frac{1}{n}\sum_{i=1}^{n-1}(n-i)a_i c_{n-i}. \quad (3.12)$$

Число кепстральных коэффициентов необязательно должно быть равно числу коэффициентов фильтра. Преимущество кепстральных коэффициентов в том, что они, как правило, декоррелированы. Кроме того, кепстр позволяет разделить передаточную функцию передающего тракта и речь, поскольку они перемножаются в спектральной области и после логарифмирования оказываются аддитивными. Однако существует небольшая проблема в том, что старшие коэффициенты кепстра имеют малые числовые значения, что может быть компенсировано введением весовых коэффициентов в соответствующую метрику сравнения.

Кепстр сигнала на основе спектра Фурье вычисляется путем применения косинусного Фурье преобразования к логарифму спектра:

$$c_j = \sqrt{\frac{2}{N}}\sum_{i=0}^{N-1}[s_i \cos(\frac{\pi(j+1)(i+0.5)}{N})] = \sum_{i=0}^{N-1} C_{j,i} s_i, \quad (3.13)$$

где s_i – логарифм спектра, N – количество отсчётов спектра, $C_{i,j}$ – унитарная матрица косинусного преобразования.

Кепстральные коэффициенты, полученные приведённым способом из мел спектра Фурье (рис. 3.1), широко используются для распознавания с помощью марковских моделей (см. следующие разделы) и носят название MFCC (Mel-frequency cepstral coefficients).

Для того чтобы учесть динамику процесса, кепстральные коэффициенты дополняют дельта и дельта-дельта признаками, имеющими смысл дифференциала, коррелирующего со скоростью изменения признака. Дельта признак для некоторой кепстральной «полосы» с номером j вычисляется следующим образом:

$$d_j(t) = \sum_{i=-k}^{k} i c_j(t+i), \quad (3.14)$$

где t – момент времени, k обычно принимает значение 2 или 3.

Аналогичная операция, применённая к дельта-признакам, даст дельта-дельта признаки. Нетрудно заметить, что формула (3.14) описывает КИХ-фильтр (фильтр с конечной импульсной характеристикой). Амплитудно-частотная

характеристика такого фильтра для частоты следования признаков 100 Гц изображена на рис. 3.3.

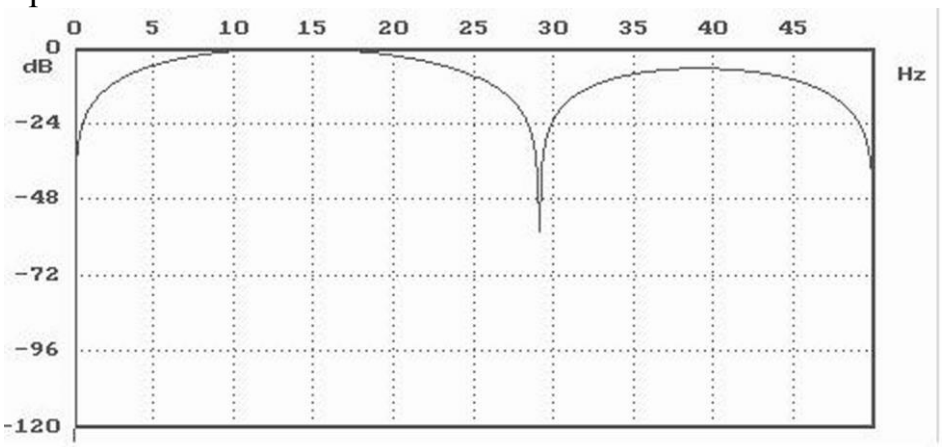


Рис. 3.3. Амплитудная характеристика дельта преобразования второго порядка

Заметим, что основными особенностями АЧХ являются:

1. устранение постоянной составляющей,
2. небольшое усиление компонент в области 10–15 Гц,
3. небольшое ослабление высокочастотных компонент.

Полезность этих особенностей понятна – устранение постоянной составляющей увеличивает динамический диапазон, устраняет влияние АЧХ канала передачи, усиление в области 10–15 Гц в некоторой степени соответствует идее усиления естественной силлабической (слоговой) частоты речи, ослабление высокочастотных компонент подавляет шумы в признаках. Явное преследование этих целей позволяет достичь лучших результатов в распознавании, что доказал Х. Германский своей системой RASTA [10].

Обычно векторы признаков имеют большие размерности, причём многие их компоненты коррелируют, то есть признаки избыточны, что приводит к дополнительным ненужным вычислениям. Для устранения этого недостатка часто используются методы понижения размерности (метод Карунена-Лоэва или метод главных компонент, факторный анализ, линейный дискриминантный анализ). Линейный дискриминантный анализ, в отличие от метода Карунена-Лоэва, в явном виде моделирует различия между классами и пытается их усилить. Факторный анализ строит признаки, основываясь на различиях между классами (используется, в основном, в задачах гуманитарного направления). Методы понижения размерности были рассмотрены в разделе «Распознавание образов».

Отметим, что процедура получения кепстра в некоторой степени декоррелирует признаки. Было замечено, что функции, определяющие линейные комбинации спектральных компонент, полученные методом Карунена-Лоэва, напоминают косинусы (по крайней мере, для первых компонент). Это и навело на мысль использовать косинусное преобразование (3.13).

4. КОЛИЧЕСТВЕННАЯ ОЦЕНКА СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ

Существуют различные по сложности и прикладному значению задачи распознавания: изолированных слов (команд); ключевых слов в потоке речи; связанной речи (тщательное проговаривание текста с паузами между словами); слитной речи (разделяют диктовку в узкой тематической области, и спонтанную речь, например, в диалоге между людьми).

Оценка системы, распознающей отдельные команды, не представляет каких-либо трудностей – количество неправильно распознанных команд делится на общее количество испытаний и получается процент ошибки. Для систем, распознающих слитную речь, ситуация не столь проста.

Задача оценки систем распознавания речи нетривиальна, так как различные алгоритмы сравниваются на ограниченных базах данных, и каждый из них имеет настраиваемые параметры, да и результаты распознавания можно интерпретировать по-разному. При этом объективное оценивание и сравнение систем распознавания речи важны как для разработчиков, так и для конечных пользователей систем.

Существует количественная методика оценки, которая применяется для сравнения и сопоставления различных систем распознавания, в ней различают такие понятия как: критерий, показатель и метод. Критерий – предмет оценки или то, что нам нужно оценить (например, точность распознавания речи, скорость, робастность к шумам и т.д.). Показатель (мера, метрика) определяет конкретное свойство, которое мы оцениваем для выбранного критерия оценки (например, процент правильно распознанных слов, время обработки сигнала, уровень максимально допустимого шума при сохранении работоспособности и т.п.). Метод – способ определения соответствующего значения для данного показателя (сравнение распознанных слов с последовательностью сказанных слов, оценка времени обработки в секундах и т.д.).

Обычно при разработке систем автоматического распознавания речи используются три разных набора данных: обучающий (“train”), отладочный (“dev”) и оценочный/тестовый (“eval”).

Обучающий набор данных (обычно это наибольшая часть речевых данных) используется только для создания и обучения моделей системы. Отладочный набор данных используется для настройки и адаптации параметров автоматической системы перед финальной стадией оценки, этот набор данных должен иметь тот же формат, что и тестовые данные. Оценочные данные содержат речевые данные, которые не использовались для обучения и настройки системы, и доступны только при финальной оценке системы. Выделают два основных критерия при оценке работы систем распознавания речи, которые далее рассмотрены детально: качество распознавания и скорость обработки [11].

4.1 Показатели оценки качества распознавания речи

Для систем автоматического распознавания речи основным показателем оценки по критерию качества является точность распознавания, которая определяется как процент правильно распознанных слов (WRR — Word Recognition Rate) или, наоборот, неправильно распознанных слов (WER — Word Error Rate). Иногда также используется показатель ошибок распознавания фраз/предложений (SER — Sentence Error Rate), который является важным в диалоговых системах, где корректировка гипотезы распознавания невозможна в отличие от задачи диктовки текста. В последнее время в качестве основного показателя точности работы систем распознавания речи используется WER (его абсолютное или относительное значение), если сравниваются различные системы распознавания речи. Поскольку с развитием речевых технологий показатель WER все более приближается к нулю, то значение улучшения WER более наглядно, чем улучшение точности распознавания слов. Метод определения WER состоит в выравнивании двух текстовых строк (первая — это результат распознавания, а вторая — запись того, что было сказано в действительности) путем алгоритма динамического программирования с вычислением расстояния Левенштейна [12]. Расстояние Левенштейна представляет собой “стоимость” редактирования текстовых данных (минимальное количество или взвешенная сумма операций редактирования [13]) для преобразования первой строки во вторую с наименьшим числом операций ручной замены (S), удаления (D) и вставки (I) слов:

$$WER = \frac{S + D + I}{T} \times 100\%, \tag{4.1}$$

где T — количество слов в распознаваемой фразе.

Также для оценки качества распознавания речи используется показатель процента корректно распознанных слов (WCR — Word Correctly Recognized), он не учитывает ошибочные вставки слов, сделанные системой:

$$WCR = \frac{H}{T} \times 100\%, \;\; H = N - D - S\,, \tag{4.2}$$

где H — количество правильно распознанных слов, а N — количество произнесенных диктором слов.

Очевидно, что WER – это интуитивно понятный и адекватный показатель качества распознавания для аналитических естественных языков (класс языков, обладающих достаточно простой морфологией и системой словообразования), в которых грамматические значения однозначно выражаются самим словом (например, английский или французский). Однако другой класс синтетических языков (например, агглютинативные языки: финский, турецкий, венгерский или флективные языки: русский, украинский, казахский и т.д.), напротив, отличается богатой морфологией и развитой системой словообразования. Такие языки могут синтезировать достаточно длинные словоформы из нескольких составных частей (морфем или слогов), которые определяют грамматические признаки. При этом в беглой речи конец слова произносится не так четко как начальная часть, что

приводит к акустической неопределенности и в среднем к более высоким значениям WER по сравнению с аналитическими языками. Кроме того, многие азиатские языки (например, китайский, корейский и т.п.) используют слоги взамен слов, а тайский и некоторые другие языки не имеют явных разделителей границ слов.

С целью оценки систем распознавания речи для синтетических языков могут дополнительно применяться и другие показатели: ошибка распознавания букв/символов (LER или CER) [14] или ошибка распознавания фонов (звуков речи) [15], ошибка распознавания слогов (SylER) [16] или ошибка распознавания морфем [17]. Кроме того, для некоторых синтетических языков (например, русского) адекватным показателем является также флективная ошибка распознавания слов (IWER — Inflectional Word Error Rate) [18], которая определяется следующим образом:

$$IWER = \frac{S_{hard} \times C_{hard} + S_{soft} \times C_{soft} + D + I}{T}; \quad C_{soft} < C_{hard}; \; C_{hard} \geq 1; \; 0 \leq C_{soft} < 1, \qquad (4.3)$$

IWER приписывает вес C_{hard} всем неверным заменам слов, которые привели к изменению лексемы слова, т.е. полного слова (количество грубых ошибок распознавания S_{hard} — замен лексем) и меньший вес C_{soft} – всем негрубым ошибкам в словах, где было неверно распознано окончание словоформы, но лексема (или основа) слова распознана правильно (количество негрубых ошибок S_{soft} — замен окончания слова).

Оценка автоматического распознавания речи по показателю WER предполагает, что все слова во входной фразе одинаково информативны и важны, однако, ясно, что в задачах, отличных от диктовки текста, например, диалоговые системы или понимание (смысла) речи, некоторые значащие слова (ключевые слова) более важны, чем остальные (функциональные слова, предлоги, заполнители и т.п.). В [19] предложено оценивать точность распознавания, используя взвешенный показатель неправильно распознанных слов (WWER — Weighted Word Error Rate), который определяется как:

$$WWER = \frac{V_S + V_D + V_I}{V_T} \times 100\%, \qquad (4.4)$$

$$V_T = \sum_{W_i} v_{W_i}, V_I = \sum_{\hat{W}_i \in I} v_{\hat{W}_i}, V_D = \sum_{W_i \in D} v_{W_i}, V_S = \sum_{s_j \in S} v_{s_j},$$

$$v_{s_j} = \max\left(\sum_{\hat{W}_i \in s_j} v_{\hat{W}_i}, \sum_{W_i \in s_j} v_{W_i} \right),$$

где v_{W_i} — вес слова W_i, которое является i-м словом во входной фразе, и $v_{\hat{W}_i}$ —вес слова $\hat{W}_i$, которое является i-м словом в гипотезе распознавания, s_j — j-й замененный фрагмент фразы (или одно слово) и v_{s_j} — вес данного сегмента s_j. Таким образом, в показателе WWER, каждое слово может иметь различный вес

(установленный экспертом или автоматически) в соответствии с его влиянием на последующее понимание смысла сказанной фразы.

Национальный институт стандартов и технологий (NIST) предложил также показатель количества неправильно распознанных слов по отношению к диктору (SAWER — Speaker Attributed Word Error Rate) для задачи стенографирования речи на совещаниях [20], в которых предполагается одновременное участие нескольких различных людей. Данная задача объединяет технологии автоматического распознавания речи и диаризации (сегментации) речи дикторов (разметке звукового сигнала на фрагменты «кто когда говорил» — “Who Spoke When”). Результатом работы объединенной системы является текстовая транскрипция входного одноканального звукового сигнала с указанием говорящего для каждого распознанного слова. SAWER определяется следующим выражением:

$$SAWER = \frac{S + D + I + SE}{T} \times 100\%, \qquad (4.5)$$

где SE — число слов (или иных языковых единиц), правильно распознанных системой распознавания речи, но с неправильным указанием диктора в ходе диаризации речи дикторов.

Нужно отметить, что разработчики и пользователи должны понимать, что процент неправильного распознавания слов речи – это, в действительности, только количественный показатель точности распознавания (количество ошибок распознавания на фразу или слово), но не вероятность распознавания (вероятность неправильного распознавания слова во фразе), потому что он не ограничивается интервалом вероятности [0; 1] и не имеет верхнего предела. Например, представим, что диктор произнес фразу, состоящую из 10 слов, но система ее полностью распознала неправильно и предложила гипотезу из 12 других слов. В этом случае, WER=120 % (S=10, I=2, H=D=0), и это означает, что показатель точности WRR отрицательный (-20%), что не имеет смысла с точки зрения теории вероятностей.

Для того чтобы обойти эту проблему, были также предложены другие показатели качества распознавания речи, в частности, ошибка распознавания соответствий (MER — Match Error Rate) и показатель потери информации для слов (WIL — Word Information Lost) [21], основанные на вычислении относительной потери информации, и определяемые следующим образом:

$$MER = \frac{S + D + I}{H + S + D + I}, \quad WIL = 1 - \frac{H^2}{T \times T_O}, \text{ если } H >> S + D + I, \qquad (4.6)$$

где T_O — число слов в гипотезе распознавания, T — число слов во входной фразе. Однако оба этих показателя на практике применяются довольно редко, так как они обычно показывают несколько меньшую точность распознавания по сравнению со стандартными показателями, что не нравится разработчикам.

Все упомянутые выше показатели имеют дело только с одной наилучшей гипотезой распознавания каждой произнесенной фразы, и совсем необязательно, что этот единственный результат распознавания окажется действительно правильным. Однако многие системы распознавания речи способны выдавать сразу несколько гипотез распознавания с наибольшими вероятностями, так называемый список N лучших гипотез (N-best list). Дополнительным показателем для оценки таких результатов является показатель ошибок распознавания слов в списке из N лучших гипотез [22], который оценивается путем выбора из N гипотез предложений, ранжированных распознавателем по уменьшению оценки правдоподобия, единственной гипотезы, дающей наименьший уровень ошибок. WER для гипотезы с минимальным уровнем ошибок для каждой входной фразы выбирается как основной результат распознавания, и процент ошибок распознавания списка N лучших гипотез вычисляется для набора из этих выбранных гипотез.

При вероятностном моделировании и распознавании речи также иногда используются доверительные интервалы для того, чтобы показать значимость результатов. При оценке автоматического распознавания речи доверительный интервал (“confidence interval”) иногда указывается вместе со средним значением WER (например, WER=18,5 ± 2,3 %). В общем случае, доверительные интервалы показывают: 1) какое значение WER мы можем ожидать при изменении набора тестовых данных; 2) насколько значимым является предложенное улучшение модели распознавания. Однако на практике доверительные интервалы WER часто оказываются довольно широкими, что объясняется высокой дикторской вариативностью и речевыми сбоями (некоторые дикторы или фразы распознаются с нулевым WER, но другие дают очень высокий уровень ошибок). Большинство производимых разработчиками улучшений в системах распознавания речи не приводят к улучшению результатов, выходящих за пределы доверительного интервала WER из-за ограниченности наборов тестовых данных, что несколько снижает значимость результатов. Однако как новые, так и базовые методы распознавания речи обычно оцениваются разработчиками на одних и тех же оценочных данных (т.е. речевые данные не являются независимыми для разных сравниваемых моделей распознавания), в этом случае при количественной оценке точности распознавания доверительные интервалы могут не рассматриваться. Но в том случае, когда некоторые модели распознавания тестируются на различных и независимых тестовых наборах, требуется вычисление доверительного интервала дополнительно к среднему значению WER [23].

4.2 Показатели оценки скорости распознавания речи

Второй важный критерий оценки систем распознавания речи — скорость обработки речи, которая особенно важна в он-лайн системах распознавания речи с использованием микрофона. Она, как правило, вычисляется с использованием меры, называемой показателем скорости (SF — Speed Factor), также известной

как показатель реального времени (RT – Real Time) [20]. Он определяется как отношение общего времени обработки, требуемого для анализа всей записанной речи к длительности исходного анализируемого аудиосигнала. Например, если 10-минутный аудиофайл обрабатывается системой распознавания ровно 5 минут, то SF=0,5 реального времени, если он обрабатывается в течение 20 минут, то тогда SF=2,0 реального времени, что значительно хуже. Скорость обработки может быть также указана в абсолютных значениях времени (например, количество минут/секунд для обработки входного сигнала), что, однако, не является наглядным.

Другим показателем скорости автоматического распознавания речи является период ожидания обработки отсчета (SPL – Sample Processing Latency). Этот показатель означает максимальное количество аудиоданных, которое алгоритм распознавания должен обработать до выдачи результата для первого отсчета сигнала.

5. МЕТОД ДИНАМИЧЕСКОГО ПРОГРАММИРОВАНИЯ ДЛЯ РАСПОЗНАВАНИЯ РЕЧИ

Для распознавания команд до сих пор иногда используют метод динамического программирования (ДП), впервые предложенный в 60-х годах прошлого века. Это объясняется простотой, быстродействием и отсутствием необходимости собирать речевую базу данных.

Неизвестная команда в виде последовательности векторов признаков сравнивается с набором эталонов, представленных в таком же виде.

Основная проблема – различный темп и нелинейность темпа произнесения.

При принятии решения руководствуются критерием минимума расстояния от неизвестного произнесения до эталона. Метод подразумевает, что эталоны, принадлежащие одной команде (одному классу), группируются в кластер, то есть в компактную группу точек в некотором пространстве, в котором существует мера близости.

Идея метода проста и допускает рассмотрение на качественном уровне. Задача состоит в том, чтобы сравнить две совокупности векторов различной длины, причем на пространстве векторов есть метрика или мера близости. Представим, что мы сравниваем эталон сам с собой: отложим векторы признаков эталона по оси X и Y. На плоскости XY на пересечении координат, соответствующих векторам i и j, построим вертикальный отрезок, равный расстоянию (степени близости) между этими векторами. Тогда на квадрате со стороной, равной количеству векторов в эталоне (N), возникнет "гористый ландшафт", симметричный относительно диагонали (0,0) (N,N), однако по диагонали будет пролегать абсолютно прямая "долина" с высотой, равной 0 (поскольку расстояние от вектора до самого себя равно 0). Если мы сравниваем два различных эталона, принадлежащих одному и тому же слову, то "картина местности" исказится, однако, если используемые признаки адекватно отражают процесс восприятия, можно надеяться, что некоторая долина по-прежнему будет пролегать по ломаной, близкой к диагонали, теперь уже прямоугольника (рис. 5.1). Метод динамического программирования позволяет сосчитать минимальную сумму высот, набираемую при движении из точки (0,0) в точку (N,M) и, если это требуется, восстановить путь, по которому эта сумма набрана. Полученную сумму обычно нормируют на количество пройденных узлов, либо на сумму длин слов или длину более короткого слова и рассматривают как расстояние между двумя произнесениями. Конечно, используемые в практических системах реализации имеют множество управляемых параметров, оптимизирующих качество распознавания и уменьшающих время счета. Рассмотренный метод позволяет в дикторозависимом варианте распознавать 100–300 слов с вероятностью 90–98%.

Для придания системе дикторонезависимых качеств, для каждого слова записывают несколько эталонов от разных дикторов (в процессе обучения добавляют эталон от нового диктора, если он не распознался). Кроме того, существуют схемы нормализации эталонов относительно дикторов, а также кластеризации дикторов.

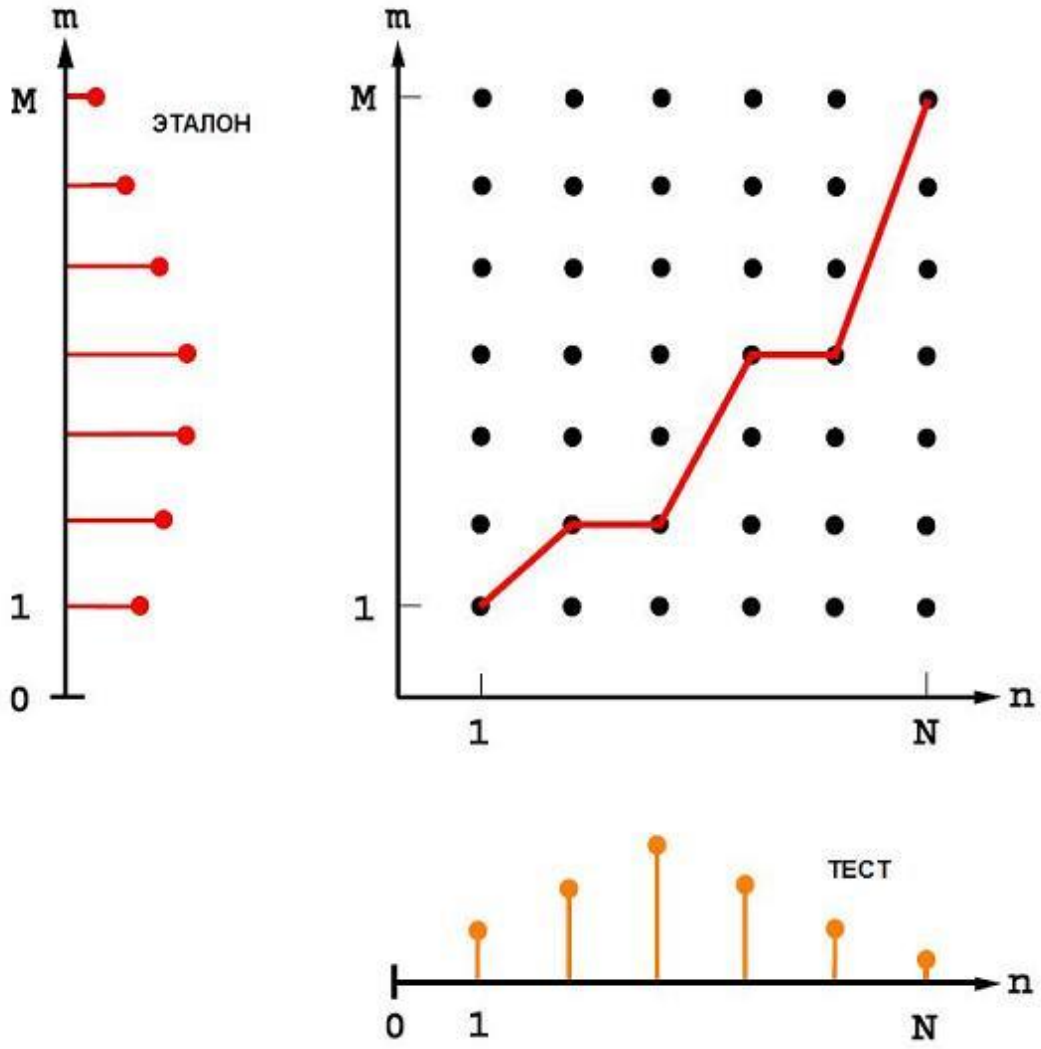


Рис. 5.1. Схема метода динамического программирования

Алгоритм распознавания команд методом ДП прозрачен и не требует подробного рассмотрения. В зависимости от топологии модели (разрешённых переходов) суммарное расстояние очередного узла матрицы [M,N] (рис. 5.1.) подсчитывается, исходя из минимума набранного расстояния:

$$S_{i,j} = Dist(i,j) + \min_{k,l \in R}(w_{k,l} S_{k,l}), \tag{5.1}$$

где $S_{n,m}$ – суммарное расстояние в узле (n,m), *Dist(i,j)* – расстояние между вектором *i* эталона и вектором *j* тестового слова, $w_{k,l}$ – вес, который присвоен узлу (*k*,*l*) относительно узла (*i*,*j*) (например, недиагональные переходы [(*i*-1,*j*)->(*i*,*j*) и (*i*,*j*-1)->(*i*,*j*)] могут «штрафоваться» бОльшим весом, чем диагональный переход [(*i*-1,*j*-1)->(*i*,*j*)]), *R* – множество разрешённых для перехода узлов (обычно это три ближайших узла [(*i*-1,*j*),(*i*-1,*j*-1) и (*i*,*j*-1)]).

Кроме суммарного расстояния, каждый узел матрицы [M,N] может содержать информацию об узле, откуда совершён переход – эта информация нужна, если требуется восстановить путь.

Оценкой близости эталона и тестового слова является нормированное суммарное расстояние правого верхнего узла матрицы. В качестве результата распознавания выбирается эталон с наименьшим расстоянием до тестового

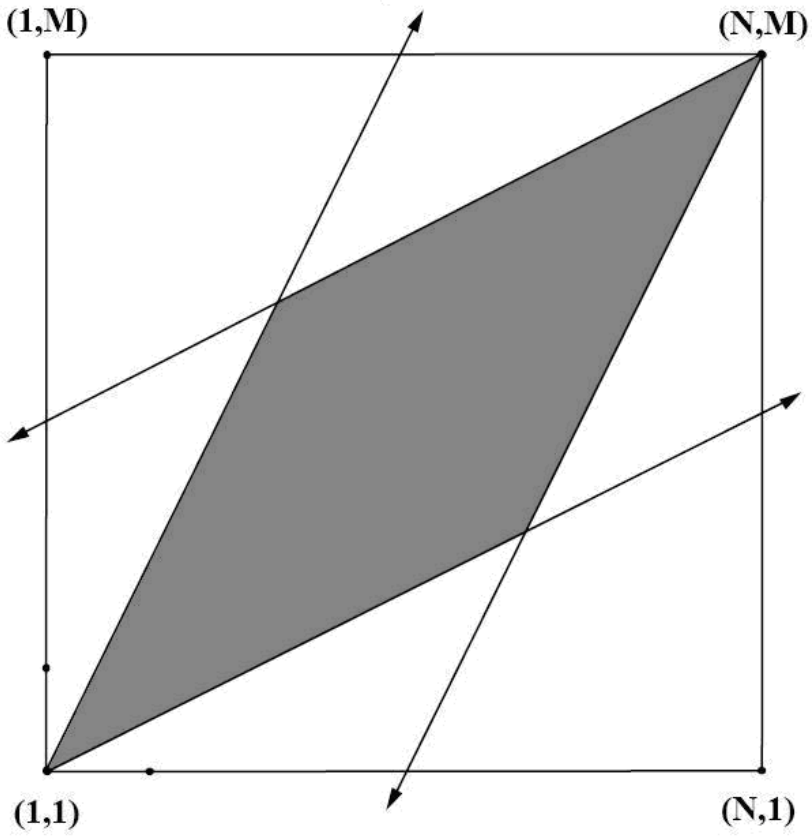


Рис. 5.2. Область подсчёта накопленных расстояний для матрицы [M,N]

слова, если расстояние не превосходит некоторого эмпирического порога, возможно, зависящего от слова.

Очевидно, что представленный алгоритм не решает задачу определения начала/конца слова. Фрагмент речи, содержащий слово, должен быть найден алгоритмом определения речь/не речь – Voice Activity Detector (VAD). Альтернативой является гораздо более трудоёмкий алгоритм ДП со скользящими концами.

Для сокращения количества вычислений можно предложить усовершенствования первоначальной модели: не рассматривать эталоны, которые отличаются по длине от тестового слова больше, чем в n раз (n обычно равно 2), не подсчитывать суммарное расстояние для узлов матрицы, далеко отстоящих от диагонали (см. рис. 5.2), прекращать вычисления, если минимальное накопленное расстояние для некоторого столбца или строки превышает порог, зависящий от номера столбца или строки.

5.1. Меры близости в пространстве признаков

В задачах распознавания речи методом ДП приходится сравнивать "похожесть" фрагментов речевого сигнала. Поскольку оценка близости

акустических событий не требует введения строгой метрики, на соблюдение второй (симметричность) и третьей (треугольника) аксиом, которым должно удовлетворять понятие метрики, в задачах распознавания речи не обращают внимания. Более того, можно представить "физиологическую" меру близости, где третья аксиома будет нарушаться закономерно, что может являться выражением того факта, что из положения артикуляторного тракта, характерного для некоторого звука, легче достичь положения, характерного для другого звука не напрямую, а через некоторый промежуточный звук. Однако чаще всего используются абстрактные меры близости, заимствованные из соответствующих разделов математики, для которых все аксиомы выполняются без всякого желания со стороны исследователя.

По-видимому, первая мера, которая использовалась в задачах распознавания – это обычная Евклидова метрика:

$$Dist(\bar{x},\bar{y}) = \sum_{i=1}^{D}(x_i - y_i)^2 \,. \tag{5.2}$$

По мере совершенствования математического аппарата, применялись меры, более адекватно отражающие характер используемых признаков. Так, использование метрики Махаланобиса:

$$Dist(\bar{x},\bar{y}) = (\bar{x}-\bar{y})^T M^{-1}(\bar{x}-\bar{y}), \tag{5.3}$$

где M – матрица ковариации, позволяло учесть корреляцию и вариативность признаков. Евклидова метрика является частным случаем метрики Махаланобиса, когда ковариационная матрица является единичной, ее имеет смысл использовать после декорреляции признаков (например, методом Карунена-Лоэва), и последующего нормирования признаков на дисперсию.

В простейших системах используют сумму модулей разностей компонент векторов признаков (квартально-блочная метрика):

$$Dist(\bar{x},\bar{y}) = \sum_{i=1}^{D} \left|x_i - y_i\right|. \tag{5.4}$$

Для кепстральных коэффициентов и коэффициентов линейного предсказания получены меры близости, отражающие характер пространств, в которых эти параметры вычисляются. Так, для кепстральных коэффициентов используется метрика Кульбака-Лейблера:

$$Dist(\bar{c}_1,\bar{c}_2) = \sum_{i=1}^{D}(c_{1,i} - c_{2,i})\ln\frac{c_{1,i}}{c_{2,i}}, \tag{5.5}$$

где суммирование проводится по всей области определения квифренси («quefrency»). Хансен показал [24], что взвешивание кепстральных коэффициентов их индексом приводит к расстоянию между наклонами спектров, а не между самими спектрами. Соответствующая мера имеет вид:

$$Dist(\bar{c}_1,\bar{c}_2) = \sum_{i=1}^{D} i(c_{1,i} - c_{2,i})^2 \,. \tag{5.6}$$

В [25] предложена "проекционная" мера:

$$Dist(\bar{c}_1,\bar{c}_2) = \left|\bar{c}_2\right|(1-\cos\beta), \tag{5.7}$$

где $\beta = \arccos(\frac{\bar{c}_2^T \bar{c}_1}{|\bar{c}_2||\bar{c}_1|})$. Показано, что эта мера более помехоустойчива.

Для коэффициентов линейного предсказания LPC используется метрика Итакуро-Саито:

$$Dist(G_1, G_2) = \int_{-\pi}^{\pi} [\left(\frac{G_1(\theta)}{G_2(\theta)}\right)^2 - \ln\left(\frac{G_1(\theta)}{G_2(\theta)}\right)^2 - 1] d\theta, \qquad (5.8)$$

где G(θ) =E/S(θ), E – энергия, S(θ) – полином, построенный на коэффициентах $\mathbf{a_i}$ (3.5–3.11), интегрирование происходит вдоль единичной окружности на Z-плоскости.

Кроме таких более или менее формальных метрик иногда вводят меры близости, основанные на данных о восприятии речи и особенностях слуховой системы. При этом оценивают расстояние между формантными структурами в шкале барк, то есть разница в частотах считается в критических полосах.

Попытки распознавать слитную речь с помощью ДП [26] утратили актуальность с развитием метода скрытых марковских моделей и поэтому далее рассматриваться не будут.

6. РАСПОЗНАВАНИЕ РЕЧИ С ПОМОЩЬЮ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ

В настоящее время большинство систем распознавания речи опираются на Скрытую Марковскую Модель (СММ). СММ – мощный статистический аппарат, представляющий спектральные свойства речи с помощью параметрического случайного процесса. Каждому моделируемому речевому объекту – фразе, слову, слогу, фонеме или аллофону (фонеме в конкретном окружении) – сопоставляется своя СММ. СММ фразы представляет собой конкатенацию СММ слов, которые представляются конкатенацией СММ более мелких элементов.

Рассмотрим математический формализм, определяющий СММ.

Процесс называется *марковским*, если для каждого момента времени вероятность любого состояния системы в следующий момент зависит только от состояния системы в настоящий момент и не зависит от того, каким образом система пришла в это состояние.

Марковский процесс называется наблюдаемым, если каждое состояние на выходе взаимно-однозначно соответствует некоторому наблюдаемому явлению.

Пример:

Состояние 1: непогода (дождь, снег, град,…)

Состояние 2: облачно

Состояние 3: солнечно

Вероятности переходов между состояниями отображены на рис.6.1 и в матрице *A* (1):

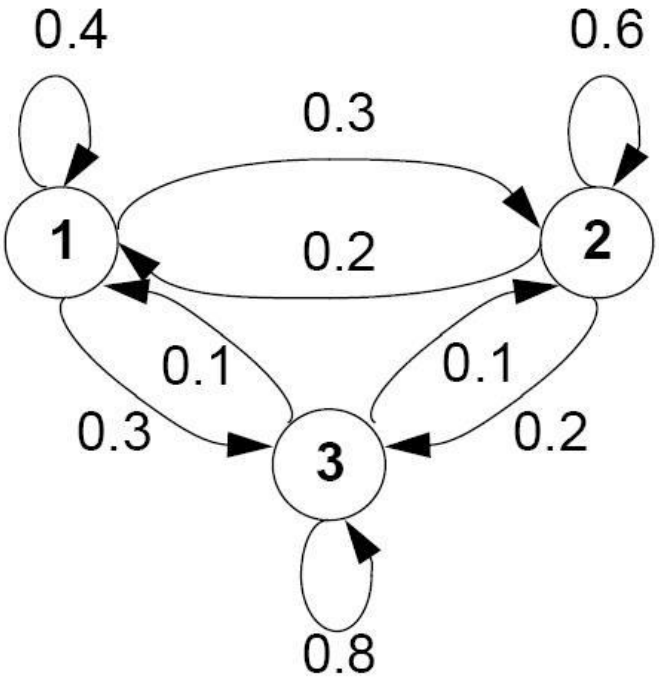


Рис.6.1. Марковская цепь, описывающая погодную модель [4]

$$A = \{a_{i,j}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}. \quad (6.1)$$

Зная исходные значения вероятностей π_j, можно рассчитать вероятность любой последовательности погодных условий в последующие дни, как произведение соответствующих вероятностей.

Если состояния связаны с наблюдаемыми явлениями вероятностным образом, то марковская цепь называется скрытой (СММ). В случае распознавания речи наблюдаемыми являются векторы признаков, которые связаны с состояниями вероятностным образом, то есть один и тот же вектор признаков может принадлежать нескольким состояниям. Таким образом, к параметрам модели добавляются распределения вероятностей состояний в пространстве признаков, которые называют вероятностями эмиссии:

$B_i(x)$ – функция плотности вероятности состояния s_i в пространстве признаков или вероятность эмиссии.

Если пространство признаков проквантовано, то **$B_i(x)$** представляется матрицей $\mathbf{B}_i(m)$, где m – номер слова в кодовой книге. Такие модели называются дискретными.

Для непрерывной модели используют аппроксимацию функции плотности вероятности набором стандартных функций – как правило, взвешенной суммой гауссовых функций. Для уменьшения количества оптимизируемых параметров используют гауссовы функции с диагональными матрицами ковариации. Существует разновидность СММ, называемая полунепрерывной – в этой СММ для аппроксимации функций плотности вероятности всех состояний используются функции из одного пула.

Задачей распознавания является сопоставление набору акустических признаков речевого сигнала или наблюдений $X(x_1,\ldots,x_n)$ последовательности слов $W(w_1,\ldots,w_k)$, имеющих наибольшую вероятность правдоподобия среди всех кандидатов:

$$W = \arg\max_W P(W|X). \quad (6.2)$$

Используя теорему Байеса, перепишем это выражение:

$$W = \arg\max_W \frac{P(W)P(X|W)}{P(X)}. \quad (6.3)$$

Поскольку в процессе распознавания вероятность уже полученных акустических признаков P(X) не подлежит оптимизации:

$$W = \arg\max_W P(W)P(X|W). \quad (6.4)$$

Вероятности (6.4) имеют простую интерпретацию: P(X|W) есть акустическая модель (вероятность порождения данной последовательности наблюдений X данной последовательностью слов W), а P(W) – вероятность существования в рассматриваемом языке данной последовательности слов (модель языка). Проблемы моделей языка будут рассмотрены в разделе 13.

Как будет показано в разделе «Выбор единиц распознавания фонетического уровня», акустическая модель представляет собой конкатенацию простых СММ, описывающих, как правило, аллофоны.

Таким образом, марковской моделью $\lambda(A,B,\pi)$ акустического события (акустической моделью), например, аллофона, называется набор из одного или нескольких состояний s_i, характеризующийся следующими параметрами:

N – количество состояний s:

π_i – начальное распределение вероятностей,

$$\sum_{i=1}^{N} \pi_i = 1.$$

$A=\{a_{i,j}\}$ – вероятность перехода из состояния s_i в состояние s_j,

$$\sum_{j=1}^{N} a_{i,j} = 1, \qquad 1 \le i \le N .$$

$B_i(x)$ или $B_i(m)$ – вероятность эмиссии:

$$\int B_i(x)dv = 1, \qquad 1 \le i \le N,$$

где интегрирование проводится по всему объёму пространства признаков,

$$\sum_{m=1}^{M} B_i(m) = 1,$$

где $1 \le i \le N$, M – размер кодовой книги, то есть количество кластеров.

Если длительность реальной последовательности наблюдений равна T, обозначим состояния в моменты времени как $(q_1,q_2,\ldots,q_T)$, тогда каждому состоянию модели s_i может соответствовать несколько последовательных значений q в соответствии с длительностью состояния. Например, $q_m=s_i$, $q_{m+1}=s_i$, $q_{m+p}=s_i$ – в данном случае состояние s_i длится p тактов.

Разрешённые переходы между состояниями в различные моменты времени определяются топологией модели, заранее заданной разработчиком. Очевидно, что переходы «назад» во времени имеет смысл запретить, то есть $a_{i,j}=0$ для $j<i$. Кроме того, имеет смысл запретить слишком дальние переходы, так как большая свобода переходов лишь увеличивает возможности для системы сделать ошибку. Обычно $a_{i,j}=0$ для $j-i>2$, то есть разрешены «петли», «переходы» и «прыжки» (под петлёй понимают переход в текущее состояние, под переходом – переход в следующее состояние, а под прыжком – переход через одно состояние). Как правило, систему ограничивают ещё больше, разрешая только петли и переходы. Это особенно удобно, поскольку, благодаря нормировке вероятностей на единицу, можно хранить не матрицу вероятностей переходов, а вектор вероятности петли $a_{i,i}$, при этом вероятность перехода будет равна $a_{i,i+1} = 1 - a_{i,i}$.

Отметим, что матрица переходов, по существу, управляет временем пребывания в состояниях и управляет неадекватно.

Зададимся вопросом: какова вероятность того, что процесс пробудет в состоянии s_i n тактов?

Попав в состояние s_i, процесс должен $n-1$ раз остаться в нём же с вероятностью $a_{i,i}$, а на такте n перейти в любое другое с вероятностью $1 - a_{i,i}$.

Таким образом:

$$P(n) = a_{i,i}^{(n-1)}(1 - a_{i,i}), \tag{6.5}$$

то есть, распределение вероятностей носит степенной характер с максимумом при нахождении в данном состоянии n=1 такт для всех состояний. Однако анализ длительностей пребывания в состояниях при обучении СММ показал, что реальное распределение вероятностей напоминает распределение Гаусса или Пуассона с максимумом при длительностях пребывания больше 1 такта (под тактом понимается период следования признаков, обычно около 0.01 сек).

Правильное моделирование временных параметров для моделей представляется достаточно важным, поскольку уменьшает возможность системе сделать ошибку, быстро пройдя состояния, не соответствующие исследуемому процессу – система будет вынуждена задержаться во всех состояниях некоторое время, значительно уменьшив вероятность идентификации ложной модели как правильной.

Учёт времени жизни состояния приводит к неоднородным или полумарковским моделям, поскольку требует нарушения принципа марковости – зависимости только от состояния процесса в предыдущий такт. Полумарковские модели будут рассмотрены в разделе 7.

Рассмотрим основные проблемы, решаемые в рамках СММ, следуя [27].

1. Проблема оценки: для данной модели $\lambda(A,B,\pi)$ и последовательности наблюдений $X(x_1,x_2,\ldots,x_T)$ вычислить вероятность $P(X|\lambda)$, то есть вероятность порождения последовательности X моделью λ. Решается алгоритмом «Вперёд-назад» (Forward-Backward).
2. Проблема распознавания: для данной последовательности наблюдений $X(x_1,x_2,\ldots,x_T)$ и модели $\lambda(A,B,\pi)$ вычислить оптимальную, в некотором смысле, последовательность состояний $Q(s_1,s_2,\ldots,s_T)$, принадлежащих модели λ. Решается алгоритмом Витерби.
3. Проблема обучения: для данной последовательности наблюдений $X(x_1,x_2,\ldots,x_T)$ и модели $\lambda(A,B,\pi)$ подстроить параметры модели так, чтобы максимизировать $P(X|\lambda)$. Решается алгоритмом Баума-Уэлша.

6.1. Алгоритм «Вперёд-Назад»

Алгоритм «Вперёд-Назад» (или «Прямого-обратного хода», Forward-Backward) включает две процедуры в соответствии со своим названием.

Рассмотрим процедуру «Вперёд».

Введём вспомогательную переменную $\alpha_t(i)$, которая представляет собой вероятность наблюдать последовательность $x_1,x_2,\ldots,x_t$ и оказаться в состоянии s_i в момент времени t для модели λ:

$$\alpha_t(i) = P(x_1, x_2, \ldots, x_t, q_t = s_i \mid \lambda). \tag{6.6}$$

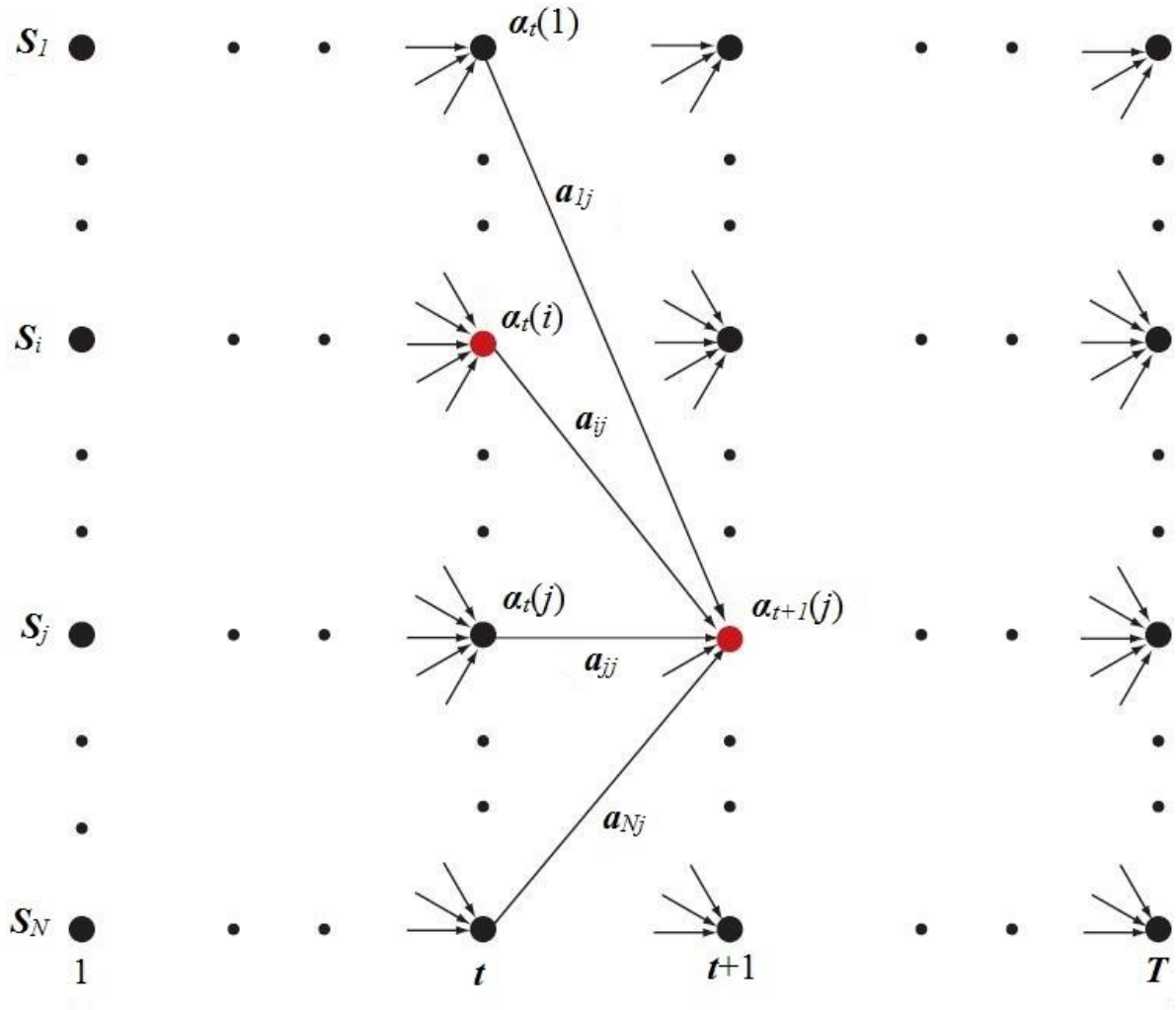


Рис. 6.2. Процедура «вперёд» (прямого хода) [2]

Рассмотрим итерационную процедуру вычисления $\alpha_t(i)$.
Инициализация:

$$\alpha_1(i) = \pi_i B_i(x_1), \ 1 \le i \le N, \tag{6.7}$$

индукция:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{i,j} \right] B_j(x_{t+1}), \ \ 1 \le t < T, 1 \le j \le N, \tag{6.8}$$

завершение:

$$P(X|\lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{6.9}$$

Объяснение данных формул достаточно очевидно:

на первом шаге вероятность определяется начальным распределением вероятностей и соответствием наблюдения x_1 данным состояниям;

на втором шаге вероятность оказаться в состоянии s_j в момент времени $t+1$ складывается из вероятностей $\alpha_t(i)$ оказаться в состоянии s_i в предыдущий момент

времени, умноженных на соответствующие вероятности переходов в состояние s_j $a_{i,j}$ с учётом соответствия наблюдения x_{t+1} состоянию s_j $B_j(x_{t+1})$ (рис. 6.2.);

на третьем шаге подсчитывается сумма всех вероятностей в конечный момент T , поскольку это независимые события.

Аналогично функционирует процедура «Назад» (рис. 6.3.).

Введём вспомогательную переменную $\beta_t(i)$, которая представляет собой вероятность наблюдать последовательность $x_{t+1},x_{t+2},\ldots,x_T$ для модели λ, при условии, что в момент t система была в состоянии s_j:

$$\beta_t(i) = P(x_{t+1}, x_{t+2},..., x_T | q_t = s_i, \lambda). \qquad (6.10)$$

Заметим, что величины $\alpha_t(i)$ и $\beta_t(i)$ дополняют друг друга в том смысле, что

$$P(X, q_t = s_i | \lambda) = \alpha_t(i)\beta_t(i), \qquad 0 \le t \le T, \qquad 1 \le i \le N. \qquad (6.11)$$

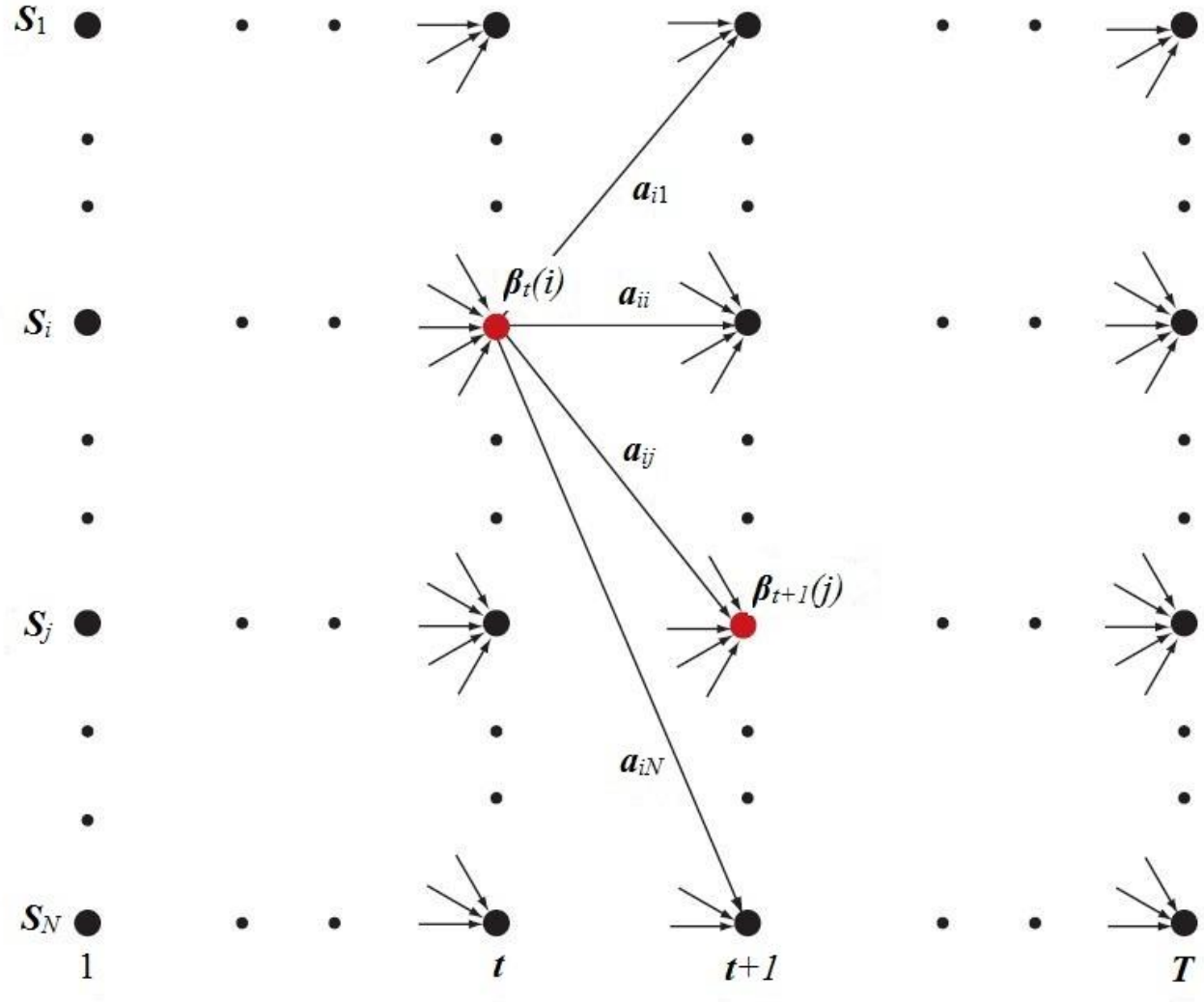


Рис. 6.3. Процедура «назад» (обратного хода) [2]

Отсюда искомая вероятность последовательности наблюдений X быть порождённой моделью λ выражается суммой вероятностей (6.11) по всем возможным состояниям для любого момента времени**:**

$$P(X|\lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i), \qquad 0 \le t \le T, \qquad 1 \le i \le N. \tag{6.12}$$

Рассмотрим итерационную процедуру вычисления $\beta_t(i)$.
Инициализация (очевидна, исходя из (6.9) и (6.12)):

$$\beta_T(i) = 1, \qquad 1 \le i \le N. \tag{6.13}$$

Индукция:

$$\beta_t(i) = \sum_{j=1}^{N} a_{i,j} B_j(x_{t+1})\beta_{t+1}(j), \quad 1 \le t \le T-1, \quad 1 \le i \le N. \tag{6.14}$$

Объяснение аналогично алгоритму «вперёд».
Завершение:

$$P(X|\lambda) = \sum_{i=1}^{N} \pi_i B_i(x_1)\beta_1(i). \tag{6.15}$$

6.2. Алгоритм Витерби

Алгоритм Витерби можно рассматривать как алгоритм динамического программирования, применённый к СММ, или как модифицированный алгоритм «вперёд», в котором вместо суммирования по всем возможным путям выбирается и запоминается наилучший путь. Определим вероятность наилучшего пути:

$$\delta_t(i) = \max_{q_1, q_2, \ldots q_{t-1}} P[q_1, q_2, \ldots, q_{t-1}, q_t = s_i, x_1, x_2, \ldots, x_t|\lambda], \tag{6.16}$$

где $\delta_t(i)$ – вероятность наиболее правдоподобной последовательности состояний, порождающей последовательность векторов признаков X_1^t и заканчивающейся в момент времени t состоянием s_i.

Для того чтобы сохранять информацию о переходах между состояниями, вводится массив $\psi_t(i)$, который содержит, например, номер состояния, из которого процесс пришёл в состояние i в момент времени t.
С введёнными обозначениями алгоритм Витерби имеет следующий вид.
Инициализация:

$$\delta_1(i) = \pi_i B_i(x_1), \qquad \psi_1(i) = 0, \quad 1 \le i \le N. \tag{6.17}$$

Индукция:

$$\delta_{t+1}(j) = \max_{1 \le i \le N}[\delta_t(i)a_{i,j}]B_j(x_{t+1}), \quad 1 \le t < T, \quad 1 \le j \le N. \tag{6.18}$$

$$\psi_t(j) = \arg\max_{1 \le i \le N}[\delta_{t-1}(i)a_{i,j}], \quad 2 \le t < T, \quad 1 \le j \le N. \tag{6.19}$$

Завершение:

$$P(X \mid \lambda) = \max_{1 \le j \le N}[\delta_T(j)] \tag{6.20}$$

$$q_T = \arg\max_{1 \le j \le N}[\delta_T(j)]. \tag{6.21}$$

Восстановление пути процесса:

$$q_t = \psi_{t+1}(q_{t+1}), t = T-1, T-2, \ldots, 1. \tag{6.22}$$

Отметим, что алгоритм Витерби получается из алгоритма «Вперёд» путём замены суммирования по состояниям на выбор максимального значения.

Поскольку вероятности, фигурирующие в формулах индукции, меньше 1, их итерационное применение приводит к потере точности. Для преодоления этой проблемы, как правило, используют логарифмы вероятностей. В этом случае все произведения превращаются в суммы, а алгоритм Витерби с вычислительной точки зрения становится тождественен классическому алгоритму динамического программирования. Отличие заключается лишь в том, что в алгоритме ДП ищется эталон с минимумом расстояния до неизвестного произнесения, а в алгоритме Витерби критерием является максимум логарифма вероятности.

6.3. Алгоритм Баума-Уэлша

Алгоритма нахождения глобального оптимума не существует – алгоритм Баума-Уэлша гарантирует нахождение локального оптимума. Для того чтобы убедиться в качестве полученного решения, можно использовать многократное обучение с различными начальными параметрами.

Введём вспомогательную величину $\xi_t(i,j)$, представляющую собой совместную вероятность находиться в состоянии s_i в момент t и в состоянии s_j в момент t+1 для данной модели λ и данной последовательности наблюдений X:

$$\xi_t(i,j) = P[q_t = s_i, q_{t+1} = s_j \,|\, X, \lambda]. \tag{6.23}$$

Используя теорему Байеса и введённые ранее параметры алгоритма Вперёд-Назад (6.6, 6.10), перепишем выражение (6.23):

$$\xi_t(i,j) = \frac{P[q_t = s_i, q_{t+1} = s_j, X \mid \lambda]}{P[X \mid \lambda]} = \frac{\alpha_t(i) a_{i,j} B_j(x_{t+1}) \beta_{t+1}(j)}{\sum_{k=1}^{N}\sum_{m=1}^{N} \alpha_t(k) a_{k,m} B_m(x_{t+1}) \beta_{t+1}(m)} \tag{6.24}$$

Введём параметр:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) = P[q_t = s_i \,|\, X, \lambda], \tag{6.25}$$

вероятность находиться в момент времени t состоянии s_i.
Учитывая индукцию (6.14):

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}. \tag{6.26}$$

$\sum_{t=1}^{T} \gamma_t(i)$ есть математическое ожидание числа попаданий в состояние s_i, а

$\sum_{t=1}^{T} \xi_t(i,j)$ – математическое ожидание числа переходов из состояния s_i в состояние s_j.

С помощью введённых величин выразим параметры модели $\lambda(A,B,\pi)$ для дискретного варианта СММ:

$$\pi_i^* = \gamma_1(i), \tag{6.27}$$

$$a^*_{i,j} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)}, \tag{6.28}$$

$$B^*_j(k) = \frac{\sum_{t=1}^{T} \gamma_t(j)[x_t = v_k]}{\sum_{t=1}^{T} \gamma_t(j)}, \tag{6.29}$$

где v_k – k-ое кодовое слово кодовой книги, $[x_t=v_k]$ обозначает, что вектор обучающей выборки x_t принадлежит кодовому слову v_k, символ '*' обозначает то, что соответствующая величина является новой оценкой.

Доказано, что оценка вероятности с новыми параметрами не хуже, чем со старыми:

$P(X|\lambda^*) >= P(X|\lambda)$.

Выведем формулы переоценки параметров вероятности эмиссии для непрерывного случая.

Представим $B_j(x)$ в виде комбинации M_j гауссовых функций:

$$B_j(X) = \sum_{m=1}^{M_j} c_{jm} G[x, \mu_{jm}, U_{jm}], \quad 1 \le m \le M_j, \tag{6.30}$$

где: X – вектор наблюдений размерности D, c_{jm} – весовой коэффициент, определяющий вклад гауссовой функции с номером m (компоненты смеси) в функцию плотности вероятности для состояния с номером j, μ_{jm} – среднее значение для компоненты смеси m состояния j, U_{jm} – ковариационная матрица компоненты смеси m состояния j,

G – многомерная гауссова функция:

$$G[X, \mu, U] = \frac{1}{(2\pi^{D/2})\sqrt{\|U\|}} \exp(-[(X-\mu)^T U^{-1}(X-\mu)]/2), \tag{6.31}$$

поскольку гауссовы функции и функция $B_j(X)$ нормированы,

$$\sum_{m=1}^{M} c_{jm} = 1, \quad 1 \le m \le M. \tag{6.32}$$

Естественно также потребовать $c_{jm}>=0$, $1<=j<=N$, $1<=m<=M$.

Модифицируем параметр $\gamma_t(i)$ (6.25) для непрерывной модели. Пусть $\gamma_t(i,m)$ – доля гауссовой компоненты с номером m в вероятности находиться в момент времени t в состоянии s_i:

$$\gamma_t(i,m) = \left[\frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}\right]\left[\frac{c_{im} G[X, \mu_{im}, U_{im}]}{\sum_{k=1}^{M} c_{ik} G[X, \mu_{ik}, U_{ik}]}\right]. \tag{6.33}$$

Формулы пересчёта параметров вероятности эмиссии для непрерывной модели имеют вид:

$$c_{ik}^{*} = \frac{\sum_{t=1}^{T} \gamma_t(i,k)}{\sum_{t=1}^{T}\sum_{m=1}^{M} \gamma_t(i,m)}, \tag{6.34}$$

$$\mu_{ik}^{*} = \frac{\sum_{t=1}^{T} \gamma_t(i,k) x_t}{\sum_{t=1}^{T} \gamma_t(i,k)}, \tag{6.35}$$

$$U_{ik}^{*} = \frac{\sum_{t=1}^{T} \gamma_t(i,k)(x_t - \mu_{ik}^{*})(x_t - \mu_{ik}^{*})^T}{\sum_{t=1}^{T} \gamma_t(i,k)}. \tag{6.36}$$

Отметим, что обучение и пересчёт параметров возможны также в рамках алгоритма Витерби. В этом случае вероятности попадания в состояние $\gamma_t(i)$ могут принимать только значения 0 или 1, поэтому в формулах (6.34−6.36) нужно использовать конкретное число векторов и сами векторы, попавшие в состояние i. Попадание в определённое состояние определяется по результатам сегментации при восстановлении пути процесса (6.22). Акустические модели, полученные методом Витерби, практически не уступают моделям, полученным методом Баума-Уэлша.

Как уже говорилось выше, ковариационные матрицы (6.36) диагонализируют для уменьшения количества оцениваемых параметров.

Для полунепрерывной модели параметры гауссовых функций μ_{jm} и U_{jm} фиксированы и пересчёту подвергаются только весовые коэффициенты c_{jm}.

Формулы пересчёта параметров для непрерывной модели (6.34–6.36) не дают ответа на вопрос: какое количество гауссовых функций в смеси следует использовать. Можно выделить три подхода к решению этого вопроса.

1. Фиксировать количество на каком-то приемлемом числе (1–20). Количество определяется объёмом обучающей выборки, вычислительными возможностями системы, сложностью конкретной прикладной задачи и т.д.

2. Эмпирически – по результатам распознавания на тестовой выборке. Остановить увеличение количества гауссовых функций в момент, когда улучшение распознавания изменяется на незначимую величину или начинает уменьшаться. Данный подход практически не применяется из-за своей трудоёмкости.

3. Использовать оценку энтропии или логарифм правдоподобия. Полную функцию правдоподобия можно оценить, используя параметр $\gamma_t(j)$ (6.26), представляющий собой вероятность в момент времени t находиться в состоянии j или вес состояния j в момент t:

$$L = \sum_{t=1}^{T} \sum_{j=1}^{N} \log(B_j(x_t)\gamma_t(j)), \tag{6.37}$$

где L – функция правдоподобия.

Для простоты рассмотрим вариант обучения с помощью алгоритма Витерби. В этом случае каждый вектор признаков обучающей выборки «приписан» только одному состоянию, поэтому не требуется проводить очень трудоёмкую операцию по одновременной оптимизации функций плотности вероятности для всех состояний. Стоит задача оценить качество аппроксимации функции плотности вероятности каждого состояния по отдельности. Логарифм правдоподобия для состояния j имеет вид:

$$L_j = \sum_{t=1}^{T} \log(B_j(x_t \delta_{t,j})), \tag{6.38}$$

где B_j определено в (6.34), символ δ_{tj} равен 1, если в момент времени t процесс находился в состоянии j и 0 в противном случае.

Нормированный на T логарифм правдоподобия (6.38) стремится к энтропии распределения с обратным знаком при $T \to \infty$ и при улучшении качества аппроксимации. Практика показывает, что при итерационном процессе, описываемом формулами (6.34–6.36), быстрый начальный рост L_j довольно скоро сменяется пологой кривой. Когда приращение L_j становится меньше некоторого порога, процесс останавливают и некоторую гауссову функцию из (6.30) (обычно функцию с наибольшим весом) расщепляют на две каким-либо способом. После этого итерационный процесс начинают снова. Когда дальнейшее увеличение количества гауссовых функций в формуле (6.30) не приводит к значимому увеличению L_j, количество гауссовых функций фиксируют. Для того чтобы избежать переобучения, можно использовать контроль аппроксимации с помощью кросс-валидационной выборки (небольшой базы данных, выделенной из обучающей выборки и не участвующей в обучении), или использовать байесовский информационный критерий.

7. НЕОДНОРОДНАЯ МАРКОВСКАЯ МОДЕЛЬ

Как было показано в предыдущем разделе, марковское требование независимости процесса от истории, следствием которого является постоянная вероятность выхода из состояния, приводит к неадекватному описанию времени жизни состояний. На рис.7.1 изображена нормированная гистограмма распределения времени пребывания в состоянии и различные варианты её аппроксимации. Преодолеть данную неадекватность можно и в рамках марковской модели. Один из применявшихся в своё время методов – расщепление состояния на несколько тождественных (с одинаковой эмиссией) состояний и запретом «прыжков» в топологии. Очевидно, что время жизни такой модели в тактах не может быть меньше, чем количество внутренних состояний. Однако также очевидно, что таким примитивным способом невозможно смоделировать реальную гистограмму.

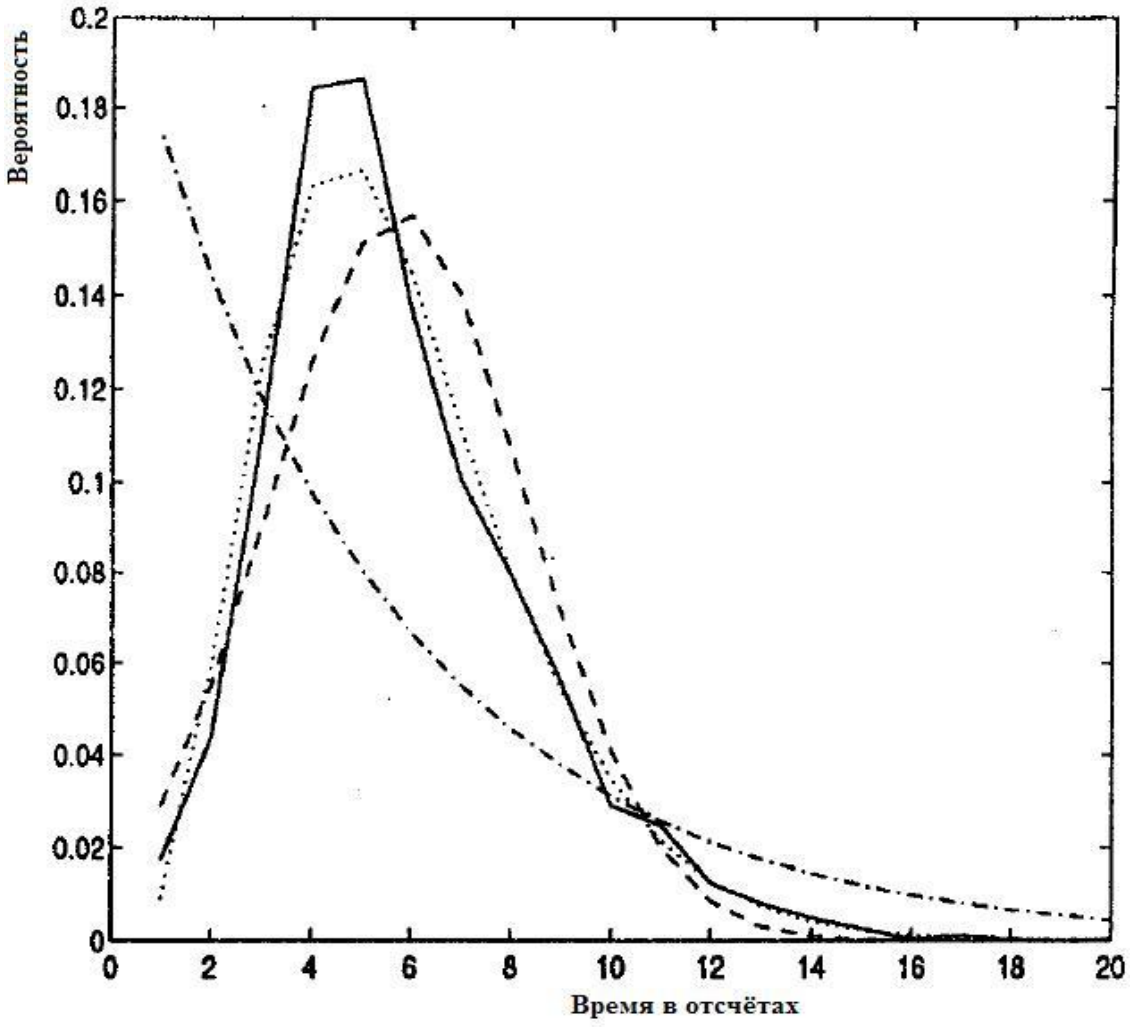


Рис.7.1. Распределение времени жизни 7-го состояния в слове «seven» [28].

Сплошная кривая – экспериментальные данные; пунктир – аппроксимация гауссовой кривой; точечная кривая – аппроксимация гамма-распределением; штрихпунктирная – распределение в соответствии с постоянной вероятностью перехода

Для того чтобы смоделировать реальную гистограмму времени жизни, приходится отказаться от предположения о независимости процесса от предыдущих состояний. В алгоритм Витерби вводится дополнительная переменная, подсчитывающая количество тактов, проведённых в каждом состоянии для каждого пути. Полученная в процессе обучения гистограмма служит источником информации о вероятности остаться в данном состоянии, как функции количества тактов, проведённых в нём – $a(t_i)$. Такая модель называется полумарковской или неоднородной марковской. Рассмотрим формулы, позволяющие по гистограмме времени пребывания состояния восстановить вероятность остаться в состоянии как функцию такта времени. Вероятность, находясь в данном состоянии, пробыть в нём n тактов, а затем покинуть его, равна:

$$P(t_n) = (1 - a(t_n))\prod_{i=1}^{n-1} a(t_i), \quad n > 1,$$
$$P(t_1) = (1 - a(t_1)). \qquad (7.1)$$

Это аналог формулы (6.5) из предыдущего раздела, где вероятность $\boldsymbol{a}$ рассматривалась как константа. Вероятности $P(t_i)$ представляют собой нормированную гистограмму.

Значения $a(t_i)$ вычисляются рекурсивно:

$$a(t_1) = 1 - P(t_1);$$
$$a(t_2) = 1 - \frac{P(t_2)}{a(t_1)};$$
$$\ldots\ldots\ldots\ldots\ldots\ldots \qquad (7.2)$$
$$a(t_n) = 1 - \frac{P(t_n)}{\prod_{i=1}^{n-1} a(t_i)};$$

В качестве $P(t_i)$ используют аппроксимацию полученной экспериментально гистограммы распределением Пуассона, нормальным или гамма-распределением. Это делается, чтобы избежать случайных скачков, связанных с недостаточностью обучающей выборки, особенно на краях гистограммы.

Рассмотренные расчёты и полученные вероятности переходов применяются к состояниям, соответствующим фонемам. При разбиении состояний, соответствующих фонемам, на подсостояния используются вероятности переходов, пересчитанные из вероятности перехода, относящиеся к фонеме в целом. Основанием для такого подхода является то, что точность оценки гистограмм времени жизни для подсостояний недостаточна.

Топология модели в виде диаграммы состояний и переходов между ними представлена на рисунке 7.2.

Прежде чем рассмотреть оценку вероятностей a_{11}, a_{22}, a_{33}, отметим, что модель рис. 7.2. не может иметь время жизни менее трёх тактов, что может не

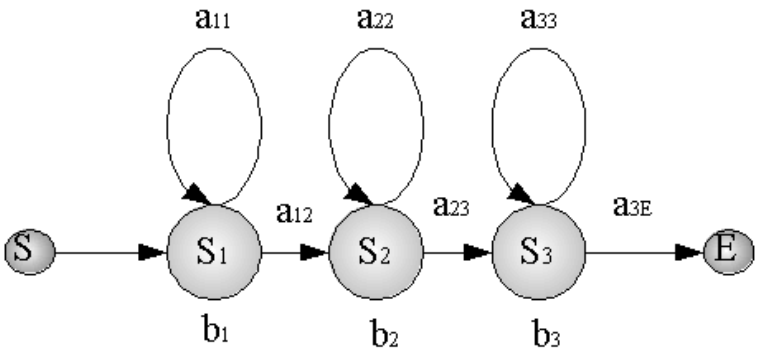


Рис. 7.2. Топология модели с тремя состояниями для трифона
S_1, S_2, S_3 – состояния модели трифона;
S, E – виртуальные узлы, соответствующие начальному (S) и конечному (E) состояниям фонемы;
a_{11}, a_{22}, a_{33} – вероятности остаться в соответствующем состоянии трифона.
a_{12}, a_{23}, – вероятности перейти в следующее состояние трифона.
a_{3e}– вероятность выйти из трифона.
b_i – плотность вероятности наблюдения вектора признаков для состояния S_i (вероятность эмиссии).

соответствовать экспериментальным данным, полученным для модели «короткоживущего» трифона. В этом случае для получения адекватного времени жизни модели трифона из трёх состояний следует изменить топологию модели и позволить прыжки из любого состояния модели в следующий трифон, минуя промежуточные состояния, однако такое решение резко увеличивает количество гипотез для декодирования. Возможным выходом будет также отказ от столь мелкого деления короткого трифона. Если же среднее время жизни трифона существенно превосходит среднеквадратичное отклонение, хорошей оценкой для вероятностей a_{11}, a_{22}, a_{33} будут вероятности, вычисленные по формулам (7.2) для нормированных гистограмм P_1, P_2, P_3, сумма средних которых равна среднему, а сумма дисперсий равна дисперсии исходной гистограммы трифона.

8. ПРОБЛЕМА ВЫБОРА ЕДИНИЦ ФОНЕТИЧЕСКОГО УРОВНЯ

В предыдущем разделе в рамках скрытой марковской модели использовалось понятие «состояние». Понятно, что для реализации распознавания речи состояния должны быть связаны с единицами фонетического уровня. Поскольку речь является процессом, возможно объединение (конкатенация) моделей фонетических фрагментов в непрерывное произнесение. Таким образом, вместо создания моделей для каждого слова, что является непосильной задачей для больших словарей, создаются модели элементов нижнего уровня. В качестве таких элементов исследовались слоги, фонемы и фрагменты фонем. В настоящее время общепринятым является использование контекстно-независимых фонем (монофонов) для средних словарей и контекстно-зависимых фонем (дифонов, трифонов) для больших словарей. Необходимость использовать части фонем и контекстную зависимость объясняется коартикуляцией (взаимным влиянием произносимых звуков друг на друга), рассмотренной в первом разделе. Взаимовлияние фонем не ограничивается соседями, а может распространяется на несколько соседних фонем. Обычно используют информацию об одном (дифоны/бифоны) или левом и правом (трифоны) соседях (по аналогии, фонемы без учёта влияния контекста называют монофонами). При этом количество фонетических единиц настолько возрастает, что даже очень больших баз данных не хватает для оценки их статистики. Приведём данные из работы [29], относящейся к английскому языку, и широко используемой базе данных Wall Street Journal Pronunciation Lexicon. Для английского языка количество фонем составляет около 50 (количество не является фиксированным – ряд распространённых дифонов (бифонов) или трифонов можно заранее отнести к отдельным фонемам). Полное количество трифонов составляет 50^3=125000. Часть этих трифонов запрещена фонетическими правилами данного языка и никогда не встречается, остаётся 95221 трифон. В упомянутой базе данных, которая составляет более 57 часов речи и содержит более 36000 предложений, встречается только 22804 трифона, из них только 14545 трифонов встречаются более 10 раз. Понятно, что для обучения СММ требуется значительное количество образцов моделируемого объекта. Число 10 можно признать минимально достаточным для обучения. Таким образом, более 80000 трифонов являются невидимыми (unseen), но могут встретиться при эксплуатации системы распознавания.

Количество параметров для одной марковской модели может достигать 1000–2000 (сюда входят матрицы переходов и параметры гауссовых функций, аппроксимирующих функции плотности вероятности). Если умножить это число на количество трифонов (около 100000), общее количество параметров, которое надо оценить в процессе обучения, оказывается порядка 10^8–10^9.Таким образом, встаёт нетривиальная задача – оценить миллионы параметров, большинство из которых в обучающей базе данных не проявляются.

Для преодоления этой трудности предложено два пути – основанный на фонетических представлениях, то есть в значительной степени субъективный, метод решающего дерева (decision tree) и процесс образования трифонов,

управляемый данными (data driven). Оба метода преследуют одну цель – управлять количеством оцениваемых параметров в зависимости от объёма обучающей выборки, поскольку опыт создания систем распознавания показал, что лучше надёжно обучить систему с небольшим количеством параметров, чем разработать сложную систему с большим количеством параметров, не обеспеченных данными для обучения.

8.1. Кластеризация на основе дерева решений

Идея метода заключается в том, что фонемы можно объединить в группы по типу влияния. Например, можно предположить, что согласные с одним и тем же местом образования одинаково влияют на последующую гласную. Тогда несколько трифонов будут описываться одной моделью. Задача метода состоит в использовании объективных критериев для объединения или дробления трифонов и бифонов.

Рассмотрим суть метода для трифонов (бифоны строятся аналогично с учётом только левого или правого контекстов).

Каждой фонеме можно приписать ряд атрибутов, последовательно конкретизирующих её свойства и разбивающих всю совокупность контекстно-независимых фонов (монофонов) на более мелкие классы. Тогда каждый трифон будет являться некоторой конечной ветвью дерева вопросов, спускаясь по которой, мы будем переходить ко всё более широким классам, объединяющим трифоны. Начиная с монофонов, будем последовательно включать информацию о правых или левых соседях, задавая относительно них «бинарные вопросы» из списка атрибутов, например «левый сосед – фрикативный? / левый сосед – не фрикативный?» (разумеется, начинать надо с вопросов, находящихся «ниже» по стволу дерева вопросов). Результатом рассматриваемой процедуры будет создание фонетического бинарного «дерева решений» (Decision tree), стволом которого является исходный монофон, а конечными ветвями или листьями – трифоны, детализированные в той степени, которую позволяют приписанные монофонам атрибуты.

В задачу построения решающего дерева входит определение последовательности вопросов, относящихся к левому и правому контекстам. Естественно стремиться к тому, чтобы очередной вопрос наилучшим, в некотором смысле, образом разделял обучающую выборку, относящуюся к данному трифону, на две выборки, относящиеся к двум уточнённым трифонам. Степень улучшения можно связать с функцией логарифма правдоподобия, получаемой при очередном делении. Функция логарифма правдоподобия системы, описываемой плотностью вероятности $p(x)$, где x – координата в пространстве признаков, имеет вид:

$$L = \sum_{i=1}^{N} \log(p(x_i)) , \qquad (8.1)$$

где x_i – совокупность векторов, принадлежащих расщепляемому трифону, N – количество векторов.

После разделения некоторого трифона получим два новых трифона, которые описываются своими плотностями вероятности с функциями логарифма правдоподобия:

$$L_{1,2} = L_1 + L_2 = \sum_{i=1}^{N_1} \log(\, p(x_i^{(1)})) + \sum_{i=1}^{N_2} \log(\, p(x_i^{(2)}))\,, \qquad (8.2)$$

где $x_i^{(1)}$, N_1 и $x_i^{(2)}$, N_2 – совокупность и количество векторов, принадлежащих первому и второму трифону, $N_1+N_2=N$.

Если каждая функция плотности вероятности аппроксимируется одним и тем же количеством гауссовых функций (рекомендуется проводить операцию построения решающего дерева, используя одну гауссову функцию на каждое состояние), совместная система будет описываться точнее, и разница:

$$\Delta L = L_1 + L_2 - L\,, \qquad (8.3)$$

являющаяся приращением правдоподобия, будет положительной.

Таким образом, становится понятна процедура получения дерева решений.

1. Относительно исходного монофона для левого и правого контекстов последовательно задаём базовые вопросы:

«левый сосед – гласная / левый сосед – не гласная»

«правый сосед – гласная / правый сосед – не гласная»

И так далее.

В дереве вопросов все монофоны делятся на гласные и согласные, поэтому отрицательный ответ на вопрос о гласной автоматически означает согласную. Представленная форма вопросов лишь подчёркивает их бинарную форму, требующую только ответов «да» или «нет».

Каждый вопрос разбивает имеющиеся в обучающей выборке монофоны на две группы, для которых после обучения и вычисления плотностей вероятности можно вычислить логарифмы правдоподобия. Первым вопросом в дереве решений становится тот, который даёт максимальное приращение правдоподобия (8.3).

2. Относительно трифона для левого и правого контекстов последовательно задаём оставшиеся вопросы из списка для исходного монофона. В качестве очередного опять ставим вопрос, дающий максимальное приращение правдоподобия.

3. Процедура для данной ветви дерева решений прекращается, когда максимальная информация, полученная на очередном шаге 2, становится меньше некоторого эмпирического порога или когда количество образцов трифона в обучающей выборке после очередного деления становится меньше некоторого порогового значения, необходимого для оценки функции плотности вероятности (этим пороговым значением в статистике часто выбирают 10, что, конечно, слишком мало для задачи оценки функции плотности вероятности в данной многомерной задаче).

Построенное фонетическое дерево решений позволяет использовать различную степень связанности состояний и, соответственно, различное количество трифонов, удовлетворяя компромисс между точностью

распознавания и доступной памятью и быстродействием. Для невидимых трифонов следует использовать наиболее близкую ветвь дерева решений.

Процедура построения решающего дерева носит чрезвычайно общий характер и не привязана исключительно к фонетике, её цель – генерация наиболее информативных единиц для распознавания. В частности, среди расщепляющих бинарных вопросов могут быть вопросы о поле, возрасте или диалектных особенностях диктора. И если обучающая база данных содержит ответы на данные вопросы, то некоторые ветви будут содержать различные трифоны для мужчин, женщин, детей и т.д. Конечно, следует помнить, что адекватная оценка функции плотности вероятности новых трифонов требует достаточной статистики. Видимо, поэтому применение очевидного вопроса о поле диктора не приводит к однозначному улучшению распознавания, хотя по информативности этот вопрос оказывается одним из первых в дереве решений [29]. Понятно, что задавая этот вопрос, мы сразу уменьшаем базу данных для исследуемого трифона наполовину (если, конечно, база содержит одинаковый объём речи для мужчин и женщин).

Процедуру, аналогичную построению решающего дерева, мы использовали для оптимизации количества гауссовых функций в смеси (предыдущий раздел). Заметим, что более точное описание функции плотности вероятности существующего трифона в некотором смысле эквивалентно его представлению в виде двух новых трифонов. Если достигнуто разделение плотностей вероятности различных трифонов в пространстве признаков, дальнейшее уточнение может происходить любым из этих путей. Добавляя гауссовы функции в смесь и проводя обучение, мы можем остановиться, когда получаемая добавочная информация становится меньше некоторого порога.

Несмотря на применяемые формулы (8.1–8.3), представляющие объективный критерий, в основе метода лежит список вопросов, составляемый лингвистами. В этой связи уместно вспомнить высказывание Ф. Джелинека, руководившего группой распознавания речи в IBM: «Every time I fire a linguist the performance of the recognizer improves».

8.2. Управляемый данными метод построения состояний

Метод основан на процедуре обучения Витерби. Как было сказано в предыдущем разделе, в данном случае все векторы признаков обучающей выборки однозначно приписываются тому или иному состоянию, то есть производится автоматическая сегментация (выравнивание, ”forced alignment”) базы данных. При этом начальная сегментация, по крайней мере, части базы данных, производится вручную для создания моделей монофонов на старте итерационного процесса. Таким образом, все векторы признаков обучающей базы данных распределены по состояниям. Поскольку тексты обучающей базы данных известны, векторы признаков, принадлежащие всем представленным в обучающей выборке трифонам (монофонам с известными соседями слева и справа), также доступны.

Наша задача – построить объединения трифонов, то есть кластеры, таким образом, чтобы:

1. входящие в кластер трифоны были близки в соответствии с каким-либо критерием;

2. количество векторов признаков, представляющих кластер в обучающей базе, было не меньше некоторого заданного числа, чтобы обеспечить достаточную статистику для оценки функции плотности вероятности кластера.

Для кластеризации применяется кросс-энтропия, которая измеряет среднее количество информации при идентификации некоторого события, если вместо «истинного» распределения p используется распределение *b*:

$$H(p,q) = -\int_V p(x)\log(b(x))dv \qquad (8.4)$$

Оценка кросс-энтропии на основе обучающей выборки осуществляется так же, как и оценка энтропии при выборе количества гауссовых функций – в виде нормированной функции логарифма правдоподобия с обратным знаком (см. предыдущий раздел):

$$\hat{H}_j \approx -\frac{1}{N_j}\sum_{i=1}^{N_j}\log(b(x_i^j)) \quad , \qquad (8.5)$$

где N_j – количество векторов признаков, принадлежащих состоянию с номером *j*, *b(x)* – функция плотности вероятности (вероятность эмиссии) кластера.

Оценивается степень близости состояния с номером *j* с кластером, имеющим функцию плотности вероятности *b(x)*. Очевидно, что кросс-энтропия тем меньше, чем ближе данное состояние к кластеру, то есть чем «лучше» накладывается собственная функция плотности вероятности распределения состояния на функцию плотности вероятности кластера. На этом основании оценку кросс-энтропии (8.5) можно использовать в качестве меры близости при кластеризации состояний. При кластеризации все векторы признаков данного состояния переносятся в ближайший кластер. После перераспределения данных между кластерами производится новый расчёт вероятности эмиссии для каждого кластера. Итерационная процедура прекращается, когда данные перестают переходить из кластера в кластер.

Описанная процедура не даёт ответа на вопрос: на сколько подсостояний разбивать начальные состояния, полученные на основе ручной сегментации? Ответ снова можно получить, используя понятия энтропии (оцениваемой с помощью нормированной функции логарифма правдоподобия) (8.1, 8.2) и информации (оцениваемой как приращение функции логарифма правдоподобия (8.3)). Если некоторое состояние представлено в обучающей выборке достаточным количеством векторов признаков, его можно разбить на два или три состояния. Начальное разбиение произвольно (при условии, что каждое состояние описывается достаточным количеством векторов). В результате итерационной процедуры обучения будут получены уточнённые границы состояний и вычислена информация (8.3). Если каждое из состояний описывается

достаточным количеством векторов и полученная информация превышает некоторый порог, данное разбиение принимается.

Помимо моделей трифонов, обычно создают модели бифонов и монофонов.

Модели бифонов и монофонов необходимы по нескольким причинам.

Первая состоит в том, что даже при использовании большой обучающей выборки, остается вероятность того, что на этапе распознавания встретится трифон, для которого при построении моделей трифонов оказалось слишком мало данных, и соответствующая ему модель не была построена. В этом случае модель требуемого трифона должна быть заменена на акустически близкую, в качестве которой может быть использована модель бифона или монофона.

Вторая причина заключается в том, что в некоторых грамматиках (например, при поиске ключевых слов) левый или правый контекст слова может быть не определен, поэтому вместо трифона необходимо использовать модель бифона с одним фиксированным контекстом и произвольным другим контекстом.

Третья причина вытекает из необходимости оптимизации сети лексикона (см. раздел, посвященный декодеру) при распознавании с большим словарем. Оптимизация производится путем объединения фонемных узлов, имеющих одинаковые идентификаторы моделей. На каждую фонему может приходиться до нескольких сотен моделей трифонов. Поэтому, если использовать модели трифонов вместо начальных фонем слов, то при большом словаре количество точек входа в лексикон будет в десятки или в сотни раз больше, чем при использовании моделей монофонов, что при распознавании приведет к пропорциональному увеличению количества гипотез, падению производительности и снижению достоверности распознавания из-за неспособности декодера корректно обрабатывать возросшие объемы данных.

Модели монофонов строятся без учета контекстов, по всем произнесениям фонемы, присутствующим в обучающей выборке.

Модели бифонов строятся с учетом только левого или только правого контекста, второй контекст остается произвольным.

В случае, когда требуемая модель трифона не найдена и должна быть заменена моделью бифона, нужно решить, какой из двух бифонов ближе к данному трифону – бифон с фиксированным левым или правым контекстом. Для решения этой задачи, после построения моделей проводится распознавание фонограмм обучающей выборки по известным транскрипциям, с использованием моделей левых или правых бифонов и подсчетом вероятностей использования моделей каждого типа, в зависимости от контекстов. При распознавании используется тот бифон, который на обучающей выборке показал лучшие результаты.

Схема построения акустических моделей представлена на рис. 8.1.

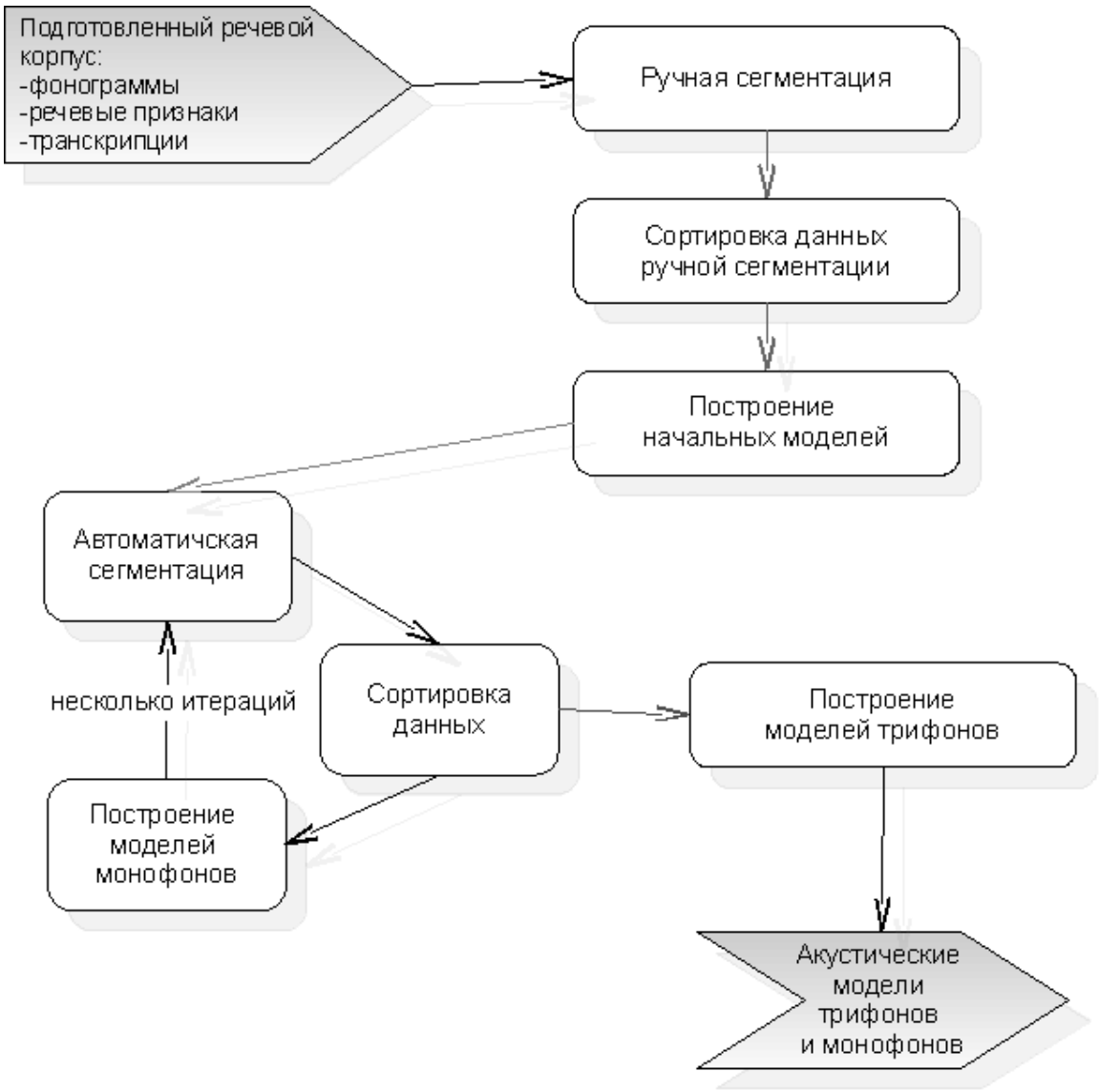


Рис. 8.1. Схема построения акустических моделей

Отметим, что при построении кластеров данным методом, в отличие от метода дерева решений, принадлежность состояний тому или иному родительскому монофону никак не учитывается. Таким образом, в один кластер могут попадать состояния от различных монофонов. Это ещё одна положительная особенность управляемого данными метода. Очевидно, что большое количество ветвей дерева решений перекрываются или находятся достаточно близко в пространстве признаков – если бы это было не так, то ФПВ исходных монофонов также не перекрывались бы в пространстве признаков, то есть распознавание монофонов выполнялось бы безошибочно, и проблема распознавания речи была решена уже на уровне монофонов, однако это не так. Поэтому метод построения дерева решений обычно дополняется процедурами «связывания состояний», смысл которых состоит в нахождении и объединении близких состояний, принадлежащих различным ветвям дерева решений. Таким образом удаётся значительно уменьшить количество независимых состояний и, следовательно, количество обучаемых параметров моделей.

9. МЕТОДЫ НОРМАЛИЗАЦИИ И АДАПТАЦИИ

Условия, в которых проходит эксплуатация систем автоматического распознавания речи, практически никогда не совпадают с условиями, в которых проходило обучение акустических моделей. Следствием этого является то, что построенные модели не являются оптимальными для данных условий. Перечислим основные факторы, искажающие речевой сигнал или обусловливающие его вариативность (рис. 9.1.):

1. Голосовой тракт и манера произнесения. Этот фактор определяет вариативность сигнала. Как бы ни была велика обучающая выборка, всегда найдутся дикторы, отличающиеся по своим характеристикам от представленных в базе.
2. Аддитивный шум, всегда присутствующий в обычных помещениях.
3. Реверберация (мультипликативный шум) – переотражённый от стен основной сигнал (рассматриваться далее не будет).
4. АЧХ микрофона и канала передачи («свёрточный шум»).
5. Аддитивный шум канала передачи.
6. Преобразование сигнала фильтром Найквиста и шум квантования (рассматриваться не будут).

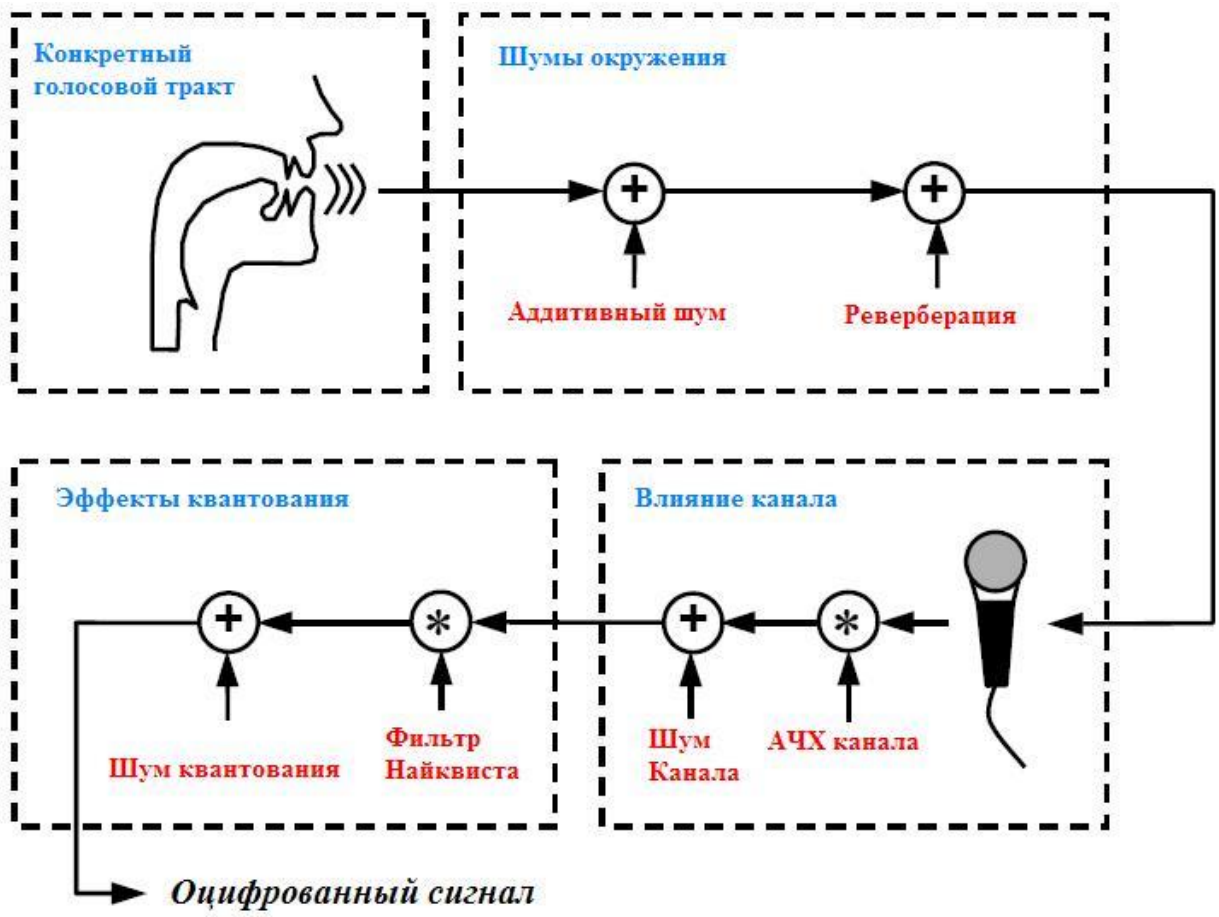


Рис.9.1. Факторы, влияющие на оцифрованный сигнал [2]

Смысл процедур нормализации и адаптации проиллюстрируем с помощью рис. 9.2. Бороться с источниками вариативности и искажений можно двумя способами: пытаться восстанавливать исходный речевой сигнал, модифицировать вычисленные признаки, или перестраивать акустические модели состояний в соответствии с возмущениями. Методы восстановления исходного сигнала из зашумлённого рассматриваются в курсе, посвящённом обработке сигналов и шумоподавлению, и поэтому здесь рассматриваться не будут. Отметим только, что обычно сигнал, прошедший процедуру шумоподавления, не обеспечивает распознавания на высоком уровне. Это происходит потому, что операции над сигналом направлены на удовлетворение субъективных критериев качества и разборчивости звука и не настроены на приведение зашумлённых признаков к признакам, характерным для чистого сигнала. Практика показывает, что гораздо лучшие результаты даёт обучение акустических моделей с помощью обучающей выборки, зашумлённой аналогичным шумом. Но обучение – это трудоёмкая процедура, а разнообразие шумов огромно, поэтому задача ставится так: используя акустические модели, полученные в каких-то фиксированных условиях (обычно с минимальным шумом), научиться распознавать речевой сигнал, полученный в других шумовых условиях.

Таким образом, в дальнейшем изложении ограничимся рассмотрением методов, преобразующих признаки, или модифицирующих акустические модели.

Методы, решающие задачу первым способом, обычно называют методами нормализации, вторым – адаптации.

Отметим, что модели гласных /i/ и /e/ дикторов HXS0 и особенно DAS1 на рис. 9.2 заметно смещены относительно дикторонезависимых моделей. Это вызовет ошибки при распознавании. Два предложенных выше метода сводятся к следующему.

Нормализация признаков – процедура преобразования признаков речевого сигнала данного диктора, благодаря которым индивидуальные модели данного диктора сместятся в области, соответствующие максимумам функций плотности вероятности (ФПВ) для соответствующих состояний дикторонезависимых моделей.

Адаптация моделей – процедура смещения дикторонезависимых моделей в сторону индивидуальных моделей данного диктора.

Способы адаптации можно разделить на адаптацию с учителем (когда известна текстовка дополнительного речевого материала нового диктора) и без учителя, а также на пакетную, инкрементную и мгновенную.

Пакетная – наиболее часто используемый вариант адаптации с учителем, когда имеется небольшая дополнительная база нового диктора.

Инкрементная – адаптация без учителя по каждой новой фразе диктора. Каждая следующая фраза распознаётся с параметрами модели, адаптированными по предыдущей.

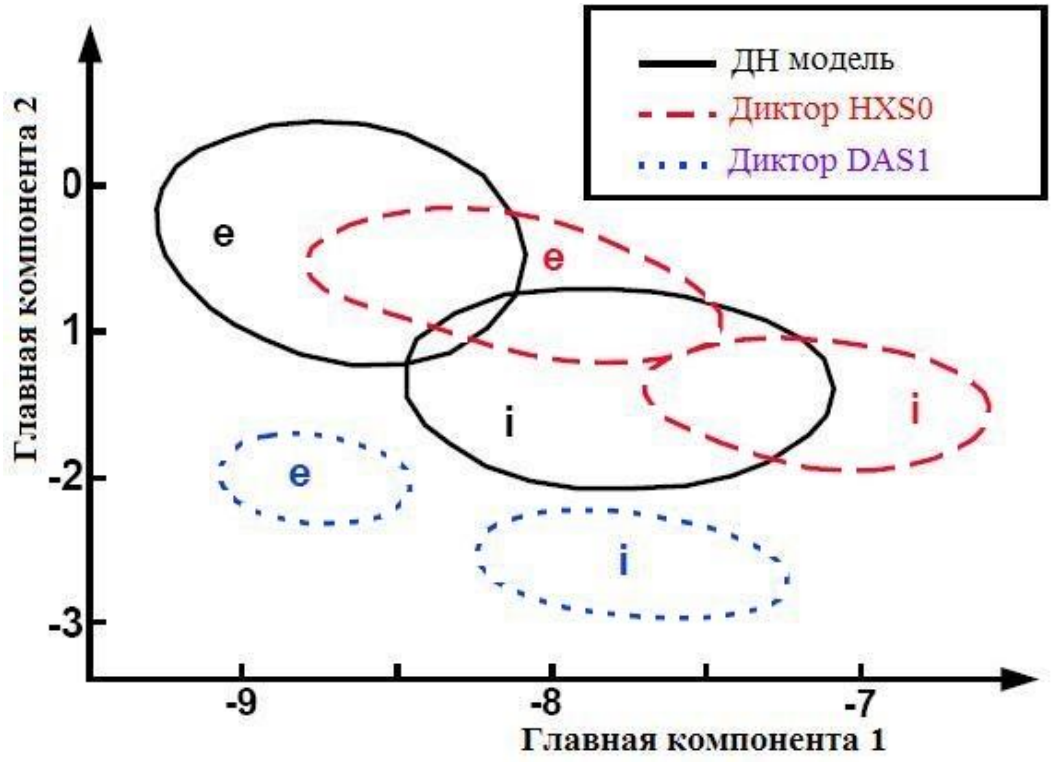


Рис. 9.2. Кривые равной плотности вероятности дикторонезависимых (ДН) и дикторозависимых моделей для монофонов /i/ и /e/ [2]

Мгновенная – по-существу, вариант инкрементной, когда речевого материала хватает только на одну адаптацию, то есть фраза распознаётся после адаптации по самой себе.

9.1. Вычитание среднего кепстра

Метод позволяет исключить АЧХ линии и микрофона, если они меняются достаточно медленно. Рассмотрим сигнал после линейного фильтра, описывающего АЧХ:

$$Y = X \otimes H, \tag{9.1}$$

где X – входной сигнал, H – импульсный отклик фильтра, Y – выходной сигнал, $\otimes$ – символ свёртки.

После преобразования Фурье выражение (9.1) принимает вид:

$$y_t = x_t h, \tag{9.2}$$

где x_t, h, y_t – векторы Фурье преобразования соответственно X, H и Y для окна, привязанного к моменту времени t. Считаем, что передаточная функция на интересующем нас интервале времени от t не зависит.

Логарифмирование приводит к выражению:

$$ly_t = lx_t + lh, \tag{9.3}$$

где lx_t, lh, ly_t – логарифмы, соответственно, x_t, h, y_t.

Применяя косинусное преобразование, получим кепстр:

$$Y_t = Cly_t = C(lx_t + lh) = X_t + H, \tag{9.4}$$

где C – матрица косинусного преобразования.

Усреднение (9.4) по некоторому интервалу времени *T* даёт:

$$\overline{Y} = \frac{1}{T}\sum_{t=0}^{T-1} Y_t = \frac{1}{T}\sum_{t=0}^{T-1}(X_t + H) = \overline{X} + H \,. \qquad (9.5)$$

Вычитая (9.5) из (9.4), получим:

$$\hat{Y}_t = Y_t - \overline{Y}_t = X_t - \overline{X}_t = \hat{X}_t, \qquad (9.6)$$

то есть избавляемся от влияния передаточной характеристики канала и микрофона.

Возникает вопрос: какова должна быть длительность усреднения T для эффективной работы алгоритма? Очевидно, что если это время сравнимо с длительностью стационарных фонем, то усреднение практически уничтожит соответствующие им векторы признаков. При слишком большой длительности алгоритм не сможет отслеживать изменения АЧХ, например, из-за движения источника звука относительно микрофона. Практика показывает, что усреднение на 2–4 секундах даёт наилучшие результаты. Метод даёт до 30% относительного улучшения качества распознавания на различных телефонных каналах. Небольшое улучшение наступает даже при применении метода в неизменном окружении с тем же микрофоном. Это можно объяснить тем, что метод «отрабатывает» движение диктора относительно микрофона, а также вычитает среднюю частотную характеристику диктора, оставляя только динамические характеристики.

Недостатком метода в описанной выше форме является то, что вычитание среднего кепстра осуществляется на всём сигнале – речи и паузах. Корректное моделирование паузы является важной составляющей метода СММ, поэтому её искажение может привести к плохим результатам. В усовершенствованном методе средний кепстр оценивается отдельно для речи и пауз. Наилучшим выделителем речи является сама система распознавания, что приводит к часто встречающейся ситуации «курица-яйцо». В качестве альтернативы двухпроходной системе распознавания предлагается, наряду с качественным детектором речи, использовать линейную комбинацию средних кепстров в пограничных зонах «речь-пауза», что позволяет уменьшить скачки кепстра при ошибках в определении границ.

Предложенная процедура получения среднего удобна при работе с файлами, однако при работе с потоком речи в реальном масштабе времени важно иметь текущую оценку среднего кепстра.

Для текущей оценки можно использовать рекурсивный фильтр первого порядка (интегратор с утечкой):

$$\begin{aligned} \overline{Y_t} &= (1-\alpha)Y_t + \alpha\overline{Y_{t-1}} \\ \hat{Y}_t &= Y_t - \overline{Y_t} = \alpha(Y_t - \overline{Y}_{t-1}), \end{aligned} \qquad (9.7)$$

где коэффициент α подбирается таким, чтобы постоянная времени фильтра τ была не меньше 5 сек. Если частоту квантования обозначить F_s, то $\alpha = 1/\exp(1/(\tau * F_s))$. Учитывая, что, $\tau * F_s >> 1$, $\alpha \approx 1 - 1/(\tau * F_s)$.

Был предложен другой фильтр [10], который, кроме «мягкого» вычитания среднего (коэффициент рекурсивной части 0.98, а не 1), обладает другими полезными свойствами (см. раздел 3):

$$\hat{Y}_t = Y_t + 0.5Y_{t-1} - 0.5Y_{t-3} - Y_{t-4} + 0.98\hat{Y}_{t-1}. \quad (9.8)$$

В статье [10] фильтр применяется к спектру, однако его свойства позволяют применять его и ему подобные фильтры непосредственно к кепстральным коэффициентам.

9.2. Адаптация акустических моделей к шуму векторными рядами Тейлора

Дальнейшее изложение следует работе [30].

Запишем в стандартной форме влияние передаточной функции («свёрточного шума») $h[m]$ и аддитивного шума n[m] на сигнал $x[m]$ (m – отсчёты времени):

$$y[m] = x[m] \otimes h[m] + n[m], \quad (9.9)$$

В частотном представлении:

$$|Y(f_i)|^2 = |X(f_i)|^2 |H(f_i)|^2 + |N(f_i)|^2, \quad f_i - \text{частота}. \quad (9.10)$$

$$\begin{aligned} &\ln|Y(f_i)|^2 = \ln(|X(f_i)|^2 |H(f_i)|^2 + |N(f_i)|^2) = \\ &\ln(|X(f_i)|^2 |H(f_i)|^2 (1 + \frac{|N(f_i)|^2}{|X(f_i)|^2 |H(f_i)|^2})) = \\ &\ln|X(f_i)|^2 + \ln|H(f_i)|^2 + \ln(1 + \exp(\ln|N(f_i)|^2 - \ln|X(f_i)|^2 - \ln|H(f_i)|^2)). \end{aligned} \quad (9.11)$$

Обозначим:

$$\begin{aligned} \mathbf{x} &= \mathbf{C}(\ln|X(f_0)|^2, |X(f_1)|^2, ..., |X(f_M)|^2), \\ \mathbf{h} &= \mathbf{C}(\ln|H(f_0)|^2, |H(f_1)|^2, ..., |H(f_M)|^2), \\ \mathbf{n} &= \mathbf{C}(\ln|N(f_0)|^2, |N(f_1)|^2, ..., |N(f_M)|^2), \\ \mathbf{y} &= \mathbf{C}(\ln|Y(f_0)|^2, |Y(f_1)|^2, ..., |Y(f_M)|^2), \end{aligned} \quad (9.12)$$

где **C –** матрица косинусного преобразования.

Тогда, применив косинусное преобразование к (9.11), получим:

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \mathbf{g(n - x - h)}, \quad (9.13)$$

где функция **g(z)** имеет вид:

$$\mathbf{g(z)} = \mathbf{C}\ln(1 + \exp(\mathbf{C}^{-1}\mathbf{z})). \quad (9.14)$$

Отметим, что, поскольку матрица косинусного преобразования унитарная, $\mathbf{C}^{-1} = \mathbf{C}^T$.

Предполагаем, что **x**, **h** и **n** распределены по Гауссу со средними μ_x, μ_h, μ_n и ковариационными матрицами Σ_x, Σ_h и Σ_n и что **x**, **h** и **n** независимы.

Найдём якобиан (9.13) по отношению к **x**, **h** и **n:**

$$\left.\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right|_{(\mu_x,\mu_h,\mu_n)} = \mathbf{I} - \mathbf{C} \bullet diag\left(\frac{\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))}{1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))}\right) \bullet \mathbf{C}^{-1} =$$

$$\mathbf{I} - \mathbf{C} \bullet diag\left(\frac{1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))-1}{1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))}\right) \bullet \mathbf{C}^{-1} =$$

$$\mathbf{I} - \mathbf{C}\mathbf{I}\mathbf{C}^{-1} + \mathbf{C} \bullet diag\left(\frac{1}{1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))}\right) \bullet \mathbf{C}^{-1} =$$

$$\mathbf{C} \bullet diag\left(\frac{1}{1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))}\right) \bullet \mathbf{C}^{-1} \equiv \mathbf{A}$$

$$\left.\frac{\partial \mathbf{y}}{\partial h}\right|_{(\mu_x,\mu_h,\mu_n)} = \mathbf{C} \bullet diag\left(\frac{1}{1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))}\right) \bullet \mathbf{C}^{-1} \equiv \mathbf{A}.$$

(9.15)

.

$$\left.\frac{\partial \mathbf{y}}{\partial n}\right|_{(\mu_x,\mu_h,\mu_n)} = \mathbf{C} \bullet diag\left(\frac{\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))}{1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))}\right) \bullet \mathbf{C}^{-1} =$$

$$\mathbf{C} \bullet diag\left(\frac{1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))-1}{1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))}\right) \bullet \mathbf{C}^{-1} =$$

$$\mathbf{C}\mathbf{I}\mathbf{C}^{-1} - \mathbf{C} \bullet diag\left(\frac{1}{1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))}\right) \bullet \mathbf{C}^{-1} =$$

$$\mathbf{I} - \mathbf{C} \bullet diag\left(\frac{1}{1+\exp(\mathbf{C}^{-1}(\mathbf{n}-\mathbf{x}-\mathbf{h}))}\right) \bullet \mathbf{C}^{-1} \equiv \mathbf{I} - \mathbf{A}.$$

,

где $diag(\mathbf{a})$ – диагональная матрица с вектором **a** на диагонали.

Теперь можно приближённо оценить **y** (9.13) в окрестности точки μ_x, μ_h, μ_n с точностью до 1-го члена ряда Тейлора:

$$\mathbf{y} = \boldsymbol{\mu}_x + \boldsymbol{\mu}_h + \mathbf{g}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x - \boldsymbol{\mu}_h) + \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}_x) + \mathbf{A}(\mathbf{h} - \boldsymbol{\mu}_h) + (\mathbf{I} - \mathbf{A})(\mathbf{n} - \boldsymbol{\mu}_n) \quad (9.16)$$

Среднее от **y** может быть получено из (9.16):

$$\boldsymbol{\mu}_y = \boldsymbol{\mu}_x + \boldsymbol{\mu}_h + \mathbf{g}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x - \boldsymbol{\mu}_h). \quad (9.17)$$

Ковариация:

$$\boldsymbol{\Sigma}_y = E((\mathbf{y} - \boldsymbol{\mu}_y)(\mathbf{y} - \boldsymbol{\mu}_y)^T) =$$

$$\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)^T \mathbf{A}^T + \mathbf{A}(\mathbf{h} - \boldsymbol{\mu}_h)(\mathbf{h} - \boldsymbol{\mu}_h)^T \mathbf{A}^T + (\mathbf{I} - \mathbf{A})(\mathbf{n} - \boldsymbol{\mu}_n)(\mathbf{n} - \boldsymbol{\mu}_n)^T (\mathbf{I} - \mathbf{A})^T = \quad (9.18)$$

$$\mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T + \mathbf{A}\boldsymbol{\Sigma}_\mathbf{h}\mathbf{A}^T + (\mathbf{I} - \mathbf{A})\boldsymbol{\Sigma}_\mathbf{n}(\mathbf{I} - \mathbf{A})^T.$$

Если **h** – неизменная передаточная характеристика, то

$$\boldsymbol{\Sigma}_y = \mathbf{A}\boldsymbol{\Sigma}_\mathbf{x}\mathbf{A}^T + (\mathbf{I} - \mathbf{A})\boldsymbol{\Sigma}_\mathbf{n}(\mathbf{I} - \mathbf{A})^T. \quad (9.19)$$

Для оценки дельта и дельта-дельта признаков используют факт, что дельта признак приближённо пропорционален производной с коэффициентом 4, а из (9.16) следует:

$$\frac{\partial \mathbf{y}}{\partial t} \approx \mathbf{A}\frac{\partial \mathbf{x}}{\partial t},$$

кроме того, здесь считаем, что аддитивный шум стационарный ($\mu_{\Delta n}$=0),

тогда:

$$\boldsymbol{\mu}_{\Delta \mathbf{y}} \approx \mathbf{A}\boldsymbol{\mu}_{\Delta \mathbf{x}} \tag{9.20}$$

и

$$\boldsymbol{\Sigma}_{\Delta \mathbf{y}} \approx \mathbf{A}\boldsymbol{\Sigma}_{\Delta \mathbf{x}}\mathbf{A}^T + (\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}_{\Delta \mathbf{n}}(\mathbf{I}-\mathbf{A})^T, \tag{9.21}$$

считаем, что $\Delta \mathbf{h} = 0$, передаточная функция постоянна.
Аналогично:

$$\boldsymbol{\mu}_{\Delta\Delta \mathbf{y}} \approx \mathbf{A}\boldsymbol{\mu}_{\Delta\Delta \mathbf{x}},$$

$$\boldsymbol{\Sigma}_{\Delta\Delta \mathbf{y}} \approx \mathbf{A}\boldsymbol{\Sigma}_{\Delta\Delta \mathbf{x}}\mathbf{A}^T + (\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}_{\Delta\Delta \mathbf{n}}(\mathbf{I}-\mathbf{A})^T. \tag{9.22}$$

$\boldsymbol{\Sigma}_{\Delta \mathbf{n}}$ оценивается, исходя из независимости $\mathbf{n}_{t+2}$ и $\mathbf{n}_{t-2}$:

$$\boldsymbol{\Sigma}_{\Delta \mathbf{n}} = 2\boldsymbol{\Sigma}_{\mathbf{n}}, \tag{9.23}$$

аналогично, $\Sigma_{\Delta^2 n}$ находится, исходя из предположения о независимости $\mathbf{n}_{\Delta t+1}$ и $\mathbf{n}_{\Delta t-1}$:

$$\Sigma_{\Delta\Delta n} = 4\boldsymbol{\Sigma}_{\mathbf{n}}. \tag{9.24}$$

Для состояния j, гауссовой функции k (в окрестности точки $\mu_{x,jk}$) матрица **A** (9.15) имеет вид:

$$\mathbf{A}_{jk} = \mathbf{C} \bullet diag\left(\frac{1}{1+\exp(\mathbf{C}^{-1}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{x,jk} - \boldsymbol{\mu}_h))}\right) \bullet \mathbf{C}^{-1}, \tag{9.25}$$

а гауссово среднее (9.16) с точностью до членов ряда Тейлора 1-го порядка (см.производные (9.18), (9.21),(9.22)):

$$\begin{aligned}
&\boldsymbol{\mu}_{y,jk} = \boldsymbol{\mu}_{x,jk} + \boldsymbol{\mu}_h + \mathbf{g}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{x,jk} - \boldsymbol{\mu}_h) \approx \\
&\boldsymbol{\mu}_{x,jk} + \boldsymbol{\mu}_{h,0} + \mathbf{g}(\boldsymbol{\mu}_{n,0} - \boldsymbol{\mu}_{x,jk} - \boldsymbol{\mu}_{h,0}) + \mathbf{A}_{jk}(\boldsymbol{\mu}_h - \boldsymbol{\mu}_{h,0}) + (\mathbf{I}-\mathbf{A}_{jk})(\boldsymbol{\mu} - \boldsymbol{\mu}_{n,0}), \\
&\boldsymbol{\mu}_{\Delta y,jk} \approx \mathbf{A}_{jk}\boldsymbol{\mu}_{\Delta x,jk}, \\
&\boldsymbol{\mu}_{\Delta\Delta y,jk} \approx \mathbf{A}_{jk}\boldsymbol{\mu}_{\Delta\Delta x,jk}, \\
&\boldsymbol{\Sigma}_{\Delta \mathbf{y},jk} \approx \mathbf{A}_{jk}\boldsymbol{\Sigma}_{\Delta \mathbf{x},jk}\mathbf{A}_{jk}^T + (\mathbf{I}-\mathbf{A}_{jk})\boldsymbol{\Sigma}_{\Delta \mathbf{n}}(\mathbf{I}-\mathbf{A}_{jk})^T, \\
&\boldsymbol{\Sigma}_{\Delta\Delta \mathbf{y},jk} \approx \mathbf{A}_{jk}\boldsymbol{\Sigma}_{\Delta\Delta \mathbf{x},jk}\mathbf{A}_{jk}^T + (\mathbf{I}-\mathbf{A}_{jk})\boldsymbol{\Sigma}_{\Delta\Delta \mathbf{n}}(\mathbf{I}-\mathbf{A}_{jk})^T
\end{aligned} \tag{9.26}$$

Ковариационная матрица $\boldsymbol{\Sigma}_{\mathbf{y},jk}$ (9.11) имеет вид:

$$\boldsymbol{\Sigma}_{y,jk} = \mathbf{A}_{jk}\boldsymbol{\Sigma}_{\mathbf{x},jk}\mathbf{A}_{jk}^T + (\mathbf{I}-\mathbf{A}_{jk})\boldsymbol{\Sigma}_{\mathbf{n}}(\mathbf{I}-\mathbf{A}_{jk})^T. \tag{9.27}$$

Для нахождения параметров шума используется ОМ алгоритм (максимизации ожидания - Expectation-maximization, EM-algorithm) со стандартной вспомогательной функцией:

$$Q(\lambda, \bar{\lambda}) = E(\log(P(y_t \mid \lambda)) = \sum_t \sum_j \sum_{k \in \Omega_j} \gamma_t(j,k,\bar{\lambda}) \cdot \log p(y_t \mid k, \lambda), \tag{9.28}$$

где $\gamma_t(j,k,\lambda)$ – вероятность гауссовой функции k для вектора y в момент времени t при используемых оптимизируемых параметрах модели λ.

При использовании алгоритма Витерби значение состояния *j* определяется моментом времени *t*: $\sum_{k\in\Omega_j}\gamma_t(j,k,\bar{\lambda})=1$,если в момент времени t система находится в состоянии j.

Для определения максимума вспомогательной функции (9.28) производные от $Q(\lambda,\bar{\lambda})$ по μ_n и μ_h приравниваются 0:

$$\sum_t\sum_j\sum_{k\in\Omega_j}\gamma_t(j,k,\bar{\lambda})\cdot(\mathbf{I}-\mathbf{A}_{jk})^T\Sigma_{y,jk}^{-1}[\mathbf{y}_t-\mu_{y,jk}]=0, \qquad (9.29)$$

$$\sum_t\sum_j\sum_{k\in\Omega_j}\gamma_t(j,k,\bar{\lambda})\cdot\boldsymbol{\Sigma}_{\mathbf{y},jk}^{-1}[\mathbf{y}_t-\boldsymbol{\mu}_{\mathbf{y},jk}]=0. \qquad (9.30)$$

Подставляя $\mu_{y,jk}$ из (9.26) в (9.29), (9.30) получаем среднее шума μ_n из (9.29) и среднее канала μ_h из (9.30):

$$\boldsymbol{\mu}_n=\boldsymbol{\mu}_{n,0}+\{\sum_t\sum_j\sum_{k\in\Omega_j}\gamma_t(j,k,\bar{\lambda})\cdot(\mathbf{I}-\mathbf{A}_{jk})^T\boldsymbol{\Sigma}_{\mathbf{y},jk}^{-1}(\mathbf{I}-\mathbf{A}_{jk})\}^{-1}\bullet$$
$$\{\sum_t\sum_j\sum_{k\in\Omega_j}\gamma_t(j,k,\bar{\lambda})\cdot(\mathbf{I}-\mathbf{A}_{jk})^T\boldsymbol{\Sigma}_{\mathbf{y},jk}^{-1}[\mathbf{y}_t-\boldsymbol{\mu}_{x,jk}-\boldsymbol{\mu}_{h,0}-\mathbf{g}(\boldsymbol{\mu}_{n,0}-\boldsymbol{\mu}_{x,jk}-\boldsymbol{\mu}_{h,0})\}, \qquad (9.31)$$

$$\boldsymbol{\mu}_h=\boldsymbol{\mu}_{h,0}+\{\sum_t\sum_j\sum_{k\in\Omega_j}\gamma_t(j,k,\bar{\lambda})\cdot\mathbf{A}_{jk}{}^T\boldsymbol{\Sigma}_{\mathbf{y},jk}^{-1}\mathbf{A}_{jk}\}^{-1}\bullet$$
$$\{\sum_t\sum_j\sum_{k\in\Omega_j}\gamma_t(j,k,\bar{\lambda})\cdot\mathbf{A}_{jk}{}^T\boldsymbol{\Sigma}_{\mathbf{y},jk}^{-1}[\mathbf{y}_t-\boldsymbol{\mu}_{x,jk}-\boldsymbol{\mu}_{h,0}-\mathbf{g}(\boldsymbol{\mu}_{n,0}-\boldsymbol{\mu}_{x,jk}-\boldsymbol{\mu}_{h,0})\}. \qquad (9.32)$$

Уравнения (9.31) и (9.32) представляют шаги итерации ОМ алгоритма.

Шаги алгоритма:

1. Положить средние канала $\mu_{h,0}$ равными 0;

2. Инициализировать средний вектор шума $\mu_{n,0}$ и диагональную ковариационную матрицу $\Sigma_{n,0}$ по первым и последним отсчётам сигнала, свободным от речи;

3. Вычислить A_{jk} (9.17) и обновить параметры по формулам (9.26), (9.27);

4. Декодировать сообщение с новыми параметрами;

5. Пересчитать параметры $\gamma_t(j,k,\lambda)$ в соответствии с алгоритмом обучения и оценить новые значения средних шума и канала (9.31) и (9.32);

6. Если сегментация и параметры не установились, перейти к п. 3.

В случае, когда аддитивный шум отсутствует, а свёрточный шум представляет собой постоянную передаточную функцию, поправки к параметрам моделей приобретают чрезвычайно простую форму. Поскольку в выражении (9.11) $N(f_i)$==0, в конечной строке присутствуют только $\ln|X(f_i)|^2+\ln|H(f_i)|^2$. Проводя все вычисления согласно формулам (9.12 и 9.15), обнаружим, что матрица ***A***, определяющая якобиан, является единичной. Легко проверить, учитывая то, что **h** – неизменная передаточная характеристика, что изменению подвергнутся только средние моделей: $\boldsymbol{\mu}_y=\boldsymbol{\mu}_x+\boldsymbol{\mu}_h$, где $\boldsymbol{\mu}_h$ – вектор,

получающийся из вектора передаточной функции после логарифмирования и косинусного преобразования (то есть преобразования MFCC).

9.3. Байесовская адаптация

Предположим, что дикторонезависимые модели построены. Это означает, что известны средние и ковариационные матрицы гауссовых функций, аппроксимирующих функции плотности вероятности состояний в пространстве признаков. Если новый диктор представлен речевым материалом, не достаточным для построения собственных дикторозависимых моделей, можно, тем не менее, использовать этот материал для коррекции дикторонезависимых моделей тех трифонов, которые достаточно представлены в базе нового диктора. Рассмотрим только преобразование средних гауссовых функций $\mu_{k,i,si}$ для метода Витерби, где k – номер состояния, i – номер гауссовой функции в смеси, si – имеет смысл «дикторонезависимый» (speaker independent).

С помощью дикторонезависимых моделей мы можем отсегментировать речевой материал на участки, соответствующие выбранным моделям. Тогда средние соответствующих гауссовых функций, описывающих состояние k для нового диктора, можно рассчитать по формулам:

$$\mu_{k,i,sd} = \frac{\sum_{x_j \in s_k} \gamma_{k,i,j} x_j}{N_{k,j}} \quad , \tag{9.33}$$

где символ sd имеет смысл «дикторозависимый» (speaker dependent), s_k – состояние k, x_j – совокупность векторов дополнительной выборки нового диктора, $\gamma_{k,i,j}$ – коэффициент, с которым вектор x_j присутствует в гауссовой функции $G_{k,i}$ смеси:

$$\gamma_{k,i,j} = \frac{c_{k,i} G_{k,i}(x_j)}{\sum_m c_{k,m} G_{k,m}(x_j)}, \tag{9.34}$$

$c_{k,i}$ – коэффициенты гауссовой смеси, $N_{k,i}$ – эффективное количество векторов признаков, присутствующее в гауссовой функции $G_{k,i}$ смеси:

$$N_{k,i} = \sum_{x_j \ni s_k} \gamma_{k,i,j} . \tag{9.35}$$

Новое, откорректированное значение среднего предлагается рассчитывать по следующей формуле:

$$\hat{\mu}_{k,i} = \frac{N_{k,i}}{N_{k,i} + \tau} \mu_{k,i,sd} + \frac{\tau}{N_{k,i} + \tau} \mu_{k,i,si} . \tag{9.36}$$

Параметр τ определяется эмпирически на обучающей выборке.

Смысл формулы (9.36) заключается в том, что, если количество данных для данной компоненты гауссовой смеси невелико, то новое значение среднего будет мало отличаться от исходного дикторонезависимого варианта, если же $N_{k,i}$ велико, то новое значение будет приближаться к дикторозависимому варианту $\mu_{k,i,sd}$. Очевидно, что формула (9.36) «разрушает» структуру функции плотности

вероятности – средние гауссовых функций смещаются на расстояния, определяемые количеством данных, что может быть случайным фактором, особенно для небольших объёмов речевого материала нового диктора. Для того чтобы обойти этот недостаток, можно поступить следующим образом: параллельно полным трифонным моделям строятся монофонные модели, аппроксимированные одной гауссовой функцией. Смещение определяется по формуле (9.36), но количество гауссовых функций в смеси равно 1. Для каждого монофона определяется смещение:

$$\Delta_k = \hat{\mu}_k - \mu_{k,sd} . \qquad (9.37)$$

Далее все средние гауссовых функций, описывающих модели трифонов, являющихся наследниками данного монофона, смещаются на Δ_k. Поскольку количество параметров моделей монофонов, аппроксимированных одной гауссовой функцией, довольно мало, этот метод не требователен к объёму дополнительного обучающего материала. Данный метод получил наименование MAP – maximum a posteriori probability.

9.4. Линейная регрессия максимума правдоподобия

Данный метод адаптации позволяет получить смещение средних гауссовых функций даже для тех трифонов, которые не представлены в дополнительной обучающей выборке нового диктора.

Новые средние гауссовых функций для некоторого множества состояний предлагается искать в виде:

$$\hat{\mu}_{k,m} = W_k \cdot \xi_{k,m}, \qquad (9.38)$$

где $\hat{\mu}_{k,m}$ – преобразованный средний вектор гауссовой функции с номером m, принадлежащей гауссовой смеси, аппроксимирующей одно из состояний, принадлежащих множеству состояний с номером k, W_k – матрица преобразования размерности [n+1,n], n – размерность векторов признаков, $\xi_{k,m}$ – расширенный исходный средний вектор:

$$\xi_{k,m} = [1, \mu_{k,m,1}, \ldots, \mu_{k,m,n}]^T = [1, \mu_{k,m}]^T . \qquad (9.39)$$

Таким образом, все средние векторы гауссовых функций, аппроксимирующих функции плотности вероятности для некоторого множества состояний, преобразуются одной матрицей преобразований.

Остаётся определить принцип, согласно которому состояния будут объединяться в одинаково преобразуемые множества и оптимальный, в некотором смысле, способ вычисления матрицы преобразований по дополнительной обучающей выборке нового диктора.

В качестве одинаково преобразуемых множеств, которые в рамках данного алгоритма называются классами регрессии, можно использовать множества состояний, которые образуются на определённом уровне построения дерева решений или кластеризации, управляемой данными (см. разделы 8.1., 8.2.). Уровень объединения трифонов, на котором следует остановиться, зависит от объёма дополнительного речевого материала нового диктора и определяется эмпирически.

В работе [29] приводятся результаты эксперимента, в котором дополнительный речевой материал составлял 40 предложений, а количество классов регрессии увеличивалось от 1 (все трифоны преобразуются одинаково) до 40. Оказалось, что минимум ошибки достигался при 15 классах регрессии, а затем начинал возрастать, что свидетельствовало о недостаточности новых данных для оценки большего количества параметров (векторов средних).

Отметим, что метод позволяет и более мелкое деление, чем описано выше. Так, отнесение гауссовых функций одного состояния к двум или более классам регрессии в соответствии с некоторой мерой близости позволило бы не только смещать функцию плотности вероятности как целое, но и менять её форму. Однако обычно стоит задача получить приемлемые модели с минимумом новых данных, которые не позволяют столь мелкое деление.

Рассмотрим задачу определения матрицы преобразования W_k.

В основе метода лежит принцип максимизации правдоподобия:

$$W_k = \max_{W_k} P(X \mid \hat{\lambda}), \tag{9.40}$$

где X – векторы признаков дополнительного речевого материала нового диктора, $\hat{\lambda}$ – адаптированные параметры моделей, полученные с помощью преобразования (9.38).

Как обычно, максимизацию выражения (9.40) выполняют с помощью ОМ алгоритма и вспомогательной функции Q [31]:

$$Q = E(\log(P(X, s \mid \hat{\lambda})) = \sum_{s \in S} \sum_{t=1}^{T} P(x_t, s \mid \lambda) \log(P(x_t, s \mid \hat{\lambda}), \tag{9.41}$$

где E – математическое ожидание, λ – текущие параметры моделей, S – все возможные цепочки состояний, возникшие при распознавании речевого материала, T – количество векторов признаков в речевом материале, x_t – вектор признаков в момент t.

Очевидно, что любое изменение параметров, определяющих ФПВ моделей, приводит к изменению сегментации (длительностей состояний), что, в свою очередь, приводит к изменениям вероятностей переходов. Однако, влияние параметров ФПВ на вероятности переходов и вероятностей переходов на значение функции Q незначительно, поэтому выражение (9.41) упрощают, отбрасывая аддитивную добавку, связанную с вероятностями переходов, которую считают константой:

$$Q = \sum_{s \in S} \sum_{t=1}^{T} P(x_t, s \mid \lambda) \log(\hat{b}_s(x_t)), \tag{9.42}$$

где $\hat{b}_s$ – адаптированная ФПВ состояния s.

В сумме по состояниям участвуют все состояния данного класса регрессии.

Напомним, что согласно концепции метода Баума-Уэлша, в каждый момент времени марковский процесс проходит через все состояния с различными вероятностями, поэтому выражение (9.42) можно переписать так:

$$Q = P(X \mid \lambda) \sum_{p=1}^{K} \sum_{t=1}^{T} \gamma_t(p) \log(\hat{b}_p(x_t)), \tag{9.43}$$

где $\gamma_t(p)$ – вероятность посетить состояние s_p в момент времени t для последовательности наблюдений X (см. формулы 6.25, 6.33), K – количество состояний.

В соответствии с формулой (6.30) представим ФПВ состояния s_p в виде суммы гауссовых функций:

$$\hat{b}_p(x_t) = \sum_{m=1}^{M_p} c_m G(x_t, \hat{\mu}_{p,m}, U_{p,m}). \tag{9.44}$$

Подстановка (9.44) в (9.43) приводит к сумме логарифмов во вспомогательной функции, что делает затруднительной максимизацию. Элегантный выход предлагает алгоритм ОМ [32]. Если представить отдельные гауссовы функции проявлением ненаблюдаемых случайных параметров Y, то, учитывая формулу (6.33) для $\gamma_t(p,m)$, формально выражение (6.43) можно представить в виде:

$$\begin{aligned} Q = const + \sum_{p=1}^{K}\sum_{m=1}^{M_k}\sum_{t=1}^{T}\gamma_t(p,m)(t)(x_t - \hat{\mu}_{p,m})^T U_{p,m}^{-1}(x_t - \hat{\mu}_{p,m}) \equiv \\ const + \sum_{p=1}^{K}\sum_{m=1}^{M_k}\sum_{t=1}^{T}\gamma_t(p,m)(t)(x_t - W_k\xi_{p,m})^T U_{p,m}^{-1}(x_t - W_k\xi_{p,m}) \end{aligned}. \tag{9.45}$$

Дифференцируя (9.45) по W_k и приравнивая производную нулю, получим матричное уравнение:

$$\sum_{p=1}^{K}\sum_{m=1}^{M_k}\sum_{t=1}^{T}\gamma_t(p,m)U_{p,m}^{-1}x_t\xi_{p,m}^T = \sum_{p=1}^{K}\sum_{m=1}^{M_k}\sum_{t=1}^{T}\gamma_t(p,m)U_{p,m}^{-1}W_k\xi_{p,m}\xi_{p,m}^T. \tag{9.46}$$

В общем виде уравнение (9.46) можно решать только численно путём последовательных итераций. Однако если прибегнуть к обычному упрощению и использовать диагональные ковариационные матрицы U, возможно аналитическое решение.

Введём матрицы:

$$Z^{(K)} = \sum_{p=1}^{K}\sum_{m=1}^{M_k}\sum_{t=1}^{T}\gamma_t(p,m)U_{p,m}^{-1}x_t\xi_{p,m}^T, \tag{9.47}$$

$$D_{p,m} = \xi_{p,m}\xi_{p,m}^T, \tag{9.48}$$

$$V_{p,m} = \sum_{t=1}^{T}\gamma_t(p,m)U_{p,m}^{-1}. \tag{9.49}$$

Теперь (9.46) можно записать в следующем виде:

$$Z = \sum_{p=1}^{K}\sum_{m=1}^{M_K}V_{p,m}W_K D_{p,m}. \tag{9.49}$$

Используя то, что D – симметричная, а V – диагональная матрицы, уравнение (9.49) можно представить в поэлементном виде:

$$z_{i,j} = \sum_{q=1}^{n+1} w_{i,q}^{(K)} g_{j,q}^{(i)}, \tag{9.50}$$

где:

$$g_{j,q}^{(i)} = \sum_{p=1}^{K}\sum_{m=1}^{M_K} v_{p,m,i,i} d_{p,m,j,q}. \tag{9.51}$$

Используя для каждой строки матрицы Z свою обратную матрицу $G^{(i)-1}$, получим формулу для построчного вычисления искомой матрицы W:

$$w_i = z_i G^{(i)-1}, \tag{9.52}$$

где w_i и z_i – векторы-строки соответствующих матриц.

Поскольку матрицы $G^{(i)-1}$ сингулярные, используется алгоритм псевдообращения.

Данный метод получил название MLLR – maximum likelihood linear regression.

9.5. Метод собственных дикторов

Рассмотрим пространство на супервекторах, образованных из средних гауссовых функций каждого диктора [33, 34]. Построим матрицу из этих векторов для обучающей выборки, содержащей речь P дикторов, для аппроксимации ФПВ состояний которых используется R гауссовых функций:

$$M = \begin{bmatrix} \overline{\mu}_{1,1} \cdots \overline{\mu}_{1,R} \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ \overline{\mu}_{P,1} \cdots \overline{\mu}_{P,R} \end{bmatrix}. \tag{9.53}$$

Применив метод главных компонент к матрице M, получим собственные векторы, которые называются «собственными дикторами». Первые главные компоненты обычно соответствуют полу диктора, громкости речи, монотонности, и т.д. Используя небольшое количество первых главных компонент, можно представить нового диктора в виде линейной комбинации собственных дикторов, исходя из максимума правдоподобия $\boldsymbol{L}$ наблюдений $\boldsymbol{X}$:

$$\overline{w} = \arg\max_{\overline{w}} L(X \mid E, \overline{w}), \tag{9.54}$$

где X – речевой материал нового диктора, E – набор собственных дикторов, w – искомый вектор коэффициентов. Максимизация, как обычно, осуществляется с помощью ОМ-алгоритма путём дифференцирования стандартной вспомогательной функции (9.28, 9.42) по искомым параметрам.

9.6. Нормализация признаков по длине голосового тракта

В разделе 1.2., посвящённом вопросам речеобразования, было показано, что при сохранении формы голосового тракта частоты формант обратно пропорциональны длине тракта. Это означает, что, например, для детей и, в среднем, для женщин все форманты должны быть смещены вверх относительно «среднего мужчины». Теоретически волноводу любой длины можно придать такую форму, что собственные значения краевой задачи волнового уравнения, которые проявляются как форманты, примут любые заданные значения. Но органы управления формой голосового тракта – нёбная занавеска, язык, нижняя

челюсть, губы – имеют ограниченные возможности по управлению, поэтому указанная тенденция отчётливо проявляется в экспериментах (рис. 9.3).

Интересно отметить, что средние различия между формантами для мужских и женских голосов зависят от языка (рис. 9.4). Это может свидетельствовать о том, что речевой аппарат человека, вообще говоря, может выставлять форманты достаточно точно, но структура языка иногда не требует этого. Этот факт ставит интересные вопросы в области восприятия речи, которые выходят за рамки курса (и на которые мы не знаем ответов). Отметим два интересных момента: только в русском языке различия между первыми формантами превышают один барк, а между вторыми – полтора. (Из раздела 2 –барк является важной единицей, полученной в экспериментах по восприятию, определяющей качественно различные для восприятия звуки). При этом различия между третьими формантами существенно меньше, что довольно удивительно, поскольку третью форманту считают наименее информативной и, следовательно, менее жёстко контролируемой.

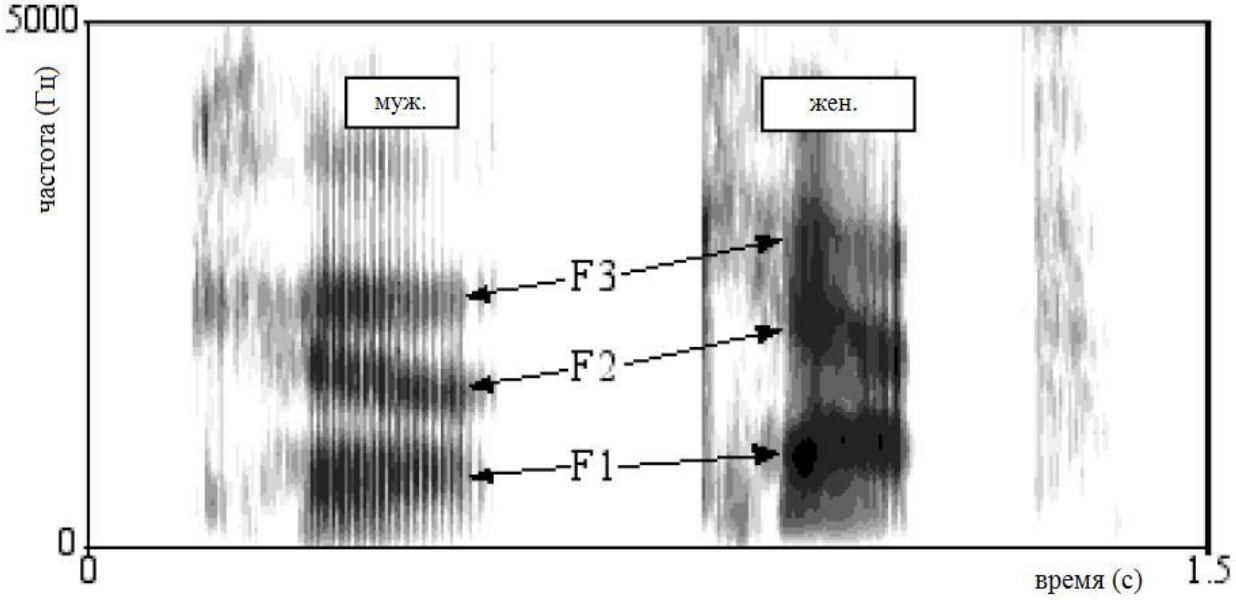


Рис. 9.3. Пример произнесения слова «cat» дикторами мужского и женского пола [36].

Если предположить, что в реальной речи значения формант действительно обратно пропорциональны длине голосового тракта, то процедура приведения формант к одинаковым значениям очевидна – достаточно линейно исказить шкалу частот. При этом даже не нужна информация о длине тракта – производится распознавание речи с различными шкалами и для дальнейшего распознавания данного диктора выбирается шкала, для которой получена наибольшая апостериорная вероятность распознавания для данного высказывания. Однако при таком подходе требуется решить две проблемы:

1. Преобразование Фурье охватывает частоты от 0 Гц до частоты Найквиста. При линейном искажении шкалы ($F_{norm}=\alpha F$) в случае $\alpha<1$ участок преобразованного спектра от αF_N до F_N не будет содержать информации, а для $\alpha>1$

исходный спектр от F_N/α до F_N не будет отражён в преобразованном (F_N – частота Найквиста).

2. Спектральный состав шумовых звуков языка определяется, в основном, параметрами сужения, а не длиной тракта.

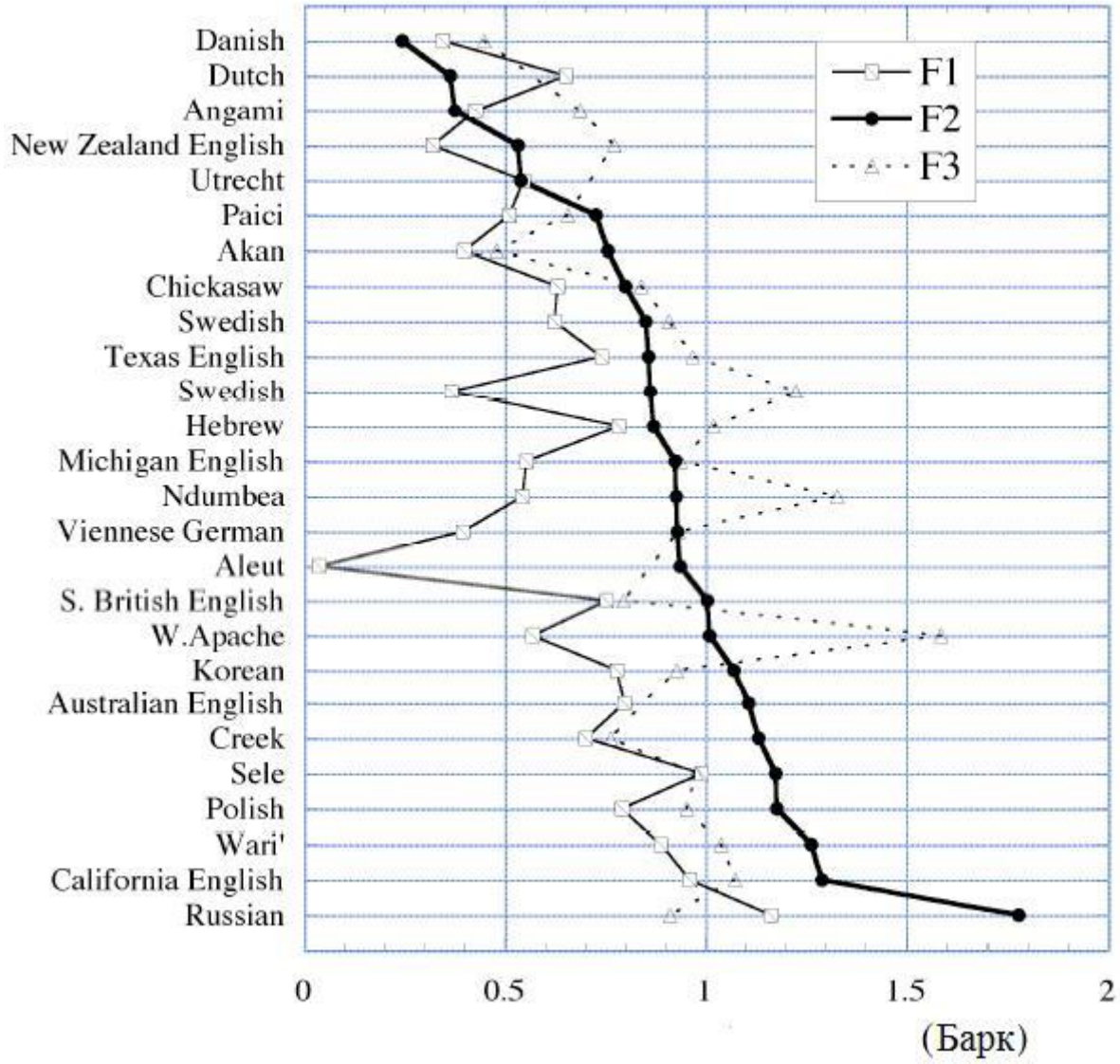


Рис. 9.4. Различия между формантами для мужских и женских голосов для различных языков [35].

Для решения указанных проблем предложены непрерывные искажающие шкалы, начинающиеся в 0 Гц и заканчивающиеся на частоте Найквиста (рис. 9.5.). Оказалось, что все они дают достаточно близкие результаты. В настоящее время предпочтение отдают шкале e). Шкалы используются следующим образом: после вычисления спектра Фурье для каждой исходной частоты спектра $f_i=i*\Delta f$, где Δf – шаг спектра ($\Delta f= F_N/W$, W – длина окна анализа), рассчитывается нормализованная частота. Значение спектра для этой частоты получают с помощью интерполяции (обычно линейной) исходного спектра. Далее нормализованный спектр используется по обычной схеме (см. MFCC признаки, раздел 3).

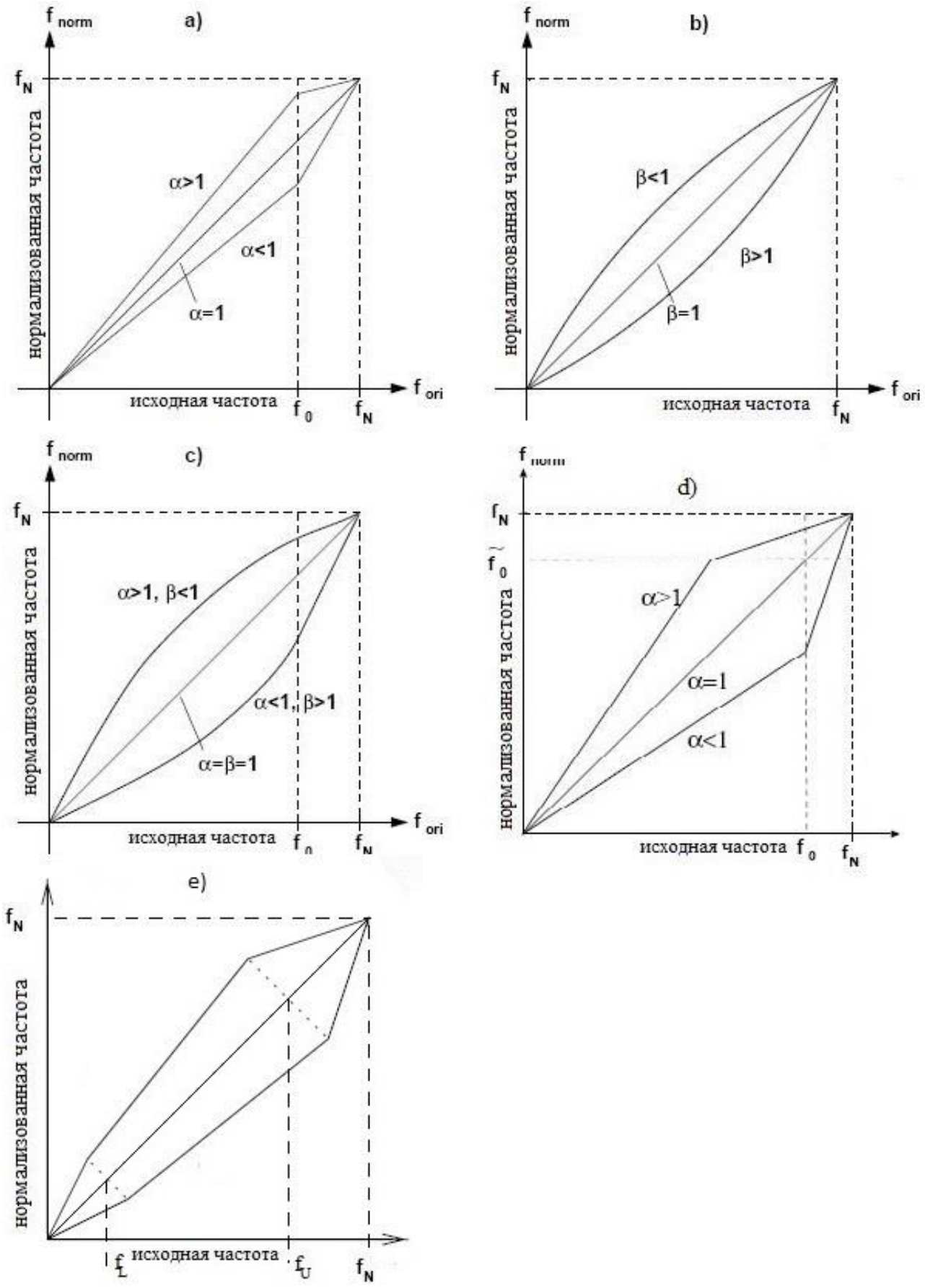


Рис. 9.5. Варианты искажающих шкал [37, 38].

Процедуру нормализации спектра можно перенести в шкалу мел. Можно также рассчитать матрицу преобразований кепстральных коэффициентов, применение которой приближённо эквивалентно искажению шкалы частот, рассмотренной выше [36]. Такой метод называется «линейной нормализацией по длине голосового тракта» (рассмотренное выше преобразование шкалы частот, очевидно, приводит к нелинейному преобразованию конечных MFCC признаков). С вычислительной точки зрения линейное преобразование более эффективно, поскольку число подлежащих интерполяции отсчётов спектра обычно равно 256, а количество кепстральных коэффициентов обычно не превышает 12–14.

Что касается проблемы невокализованных звуков, для которых форма вокального тракта не имеет большого значения, то эксперименты по раздельному нахождению коэффициентов искажения шкалы частот для различных фонем [37] практически не дали улучшения распознавания. Возможно это связано с тем, что частотные характеристики таких звуков лежат в верхнем диапазоне речевого спектра, где все шкалы сближаются.

Рассматривая методы нормализации признаков или адаптации моделей для новых дикторов, обратим внимание на то, что процедура получения моделей включает в себя процесс распознавания, то есть операцию, для более точного выполнения которой и применяются нормализация признаков и адаптация моделей. Поэтому логично эти процедуры включить в модуль обучения. Такой вид обучения называется «обучение, адаптивное к дикторам» (SAT – Speaker Adaptive Training). Ожидаемые преимущества очевидны: во-первых, модели, полученные с применением адаптивных методов компактнее в пространстве признаков и, следовательно, требуют меньше ресурсов (гауссовых функций) для своей аппроксимации, во-вторых, благодаря компактности, области, соответствующие различным моделям, имеют меньшее перекрытие, что означает меньшую ошибку распознавания (см. рис. 9.2).

10. ДИСКРИМИНАНТНЫЕ МЕТОДЫ

Рассмотрим основополагающие формулы (6.2), (6.3) из раздела 6:

$$W = \arg\max_{w} P(W \mid X), \tag{10.1}$$

$$W = \arg\max_{w} \frac{P(W)P(X \mid W)}{P(X)}, \tag{10.2}$$

где W – последовательность слов, X – последовательность векторов-признаков речевого сигнала.

Поскольку параметры модели при распознавании не изменяются, вероятность наблюдений $P(X)$ является константой. Однако в процессе обучения эта вероятность зависит от параметров моделей λ. Выразим этот факт в явном виде: $P(X|\ \lambda)$. Поскольку различные марковские модели W_j взаимно исключают друг друга,

$$P(X \mid \lambda) = \sum_{k} P(X \mid W_k, \lambda) P(W_k), \tag{10.3}$$

где суммирование производится по всем возможным последовательностям моделей. Используя (10.3), и (10.2), запишем вероятность модели W_j, явно выделяя в знаменателе соответствующий член суммы (10.3):

$$P(W_i|X,\lambda) = \frac{P(X|W_i,\lambda)P(W_i)}{P(X|W_i,\lambda)P(W_i) + \sum_{k \neq i} P(X|W_k,\lambda)P(W_k)}. \tag{10.4}$$

Обычно вероятность $P(W_i|X,\lambda)$ максимизируют в подпространстве параметров W_i, что ведёт к критерию максимального правдоподобия, то есть к оценке параметров, максимизирующих числитель (10.4). Вероятность $P(W_i)$ получают из «модели языка» (раздел 13), которая позволяет оценить вероятности последовательности слов независимо от акустического декодирования. Данный подход общеупотребителен, поскольку чрезвычайно упрощает задачу, но это достигается ценой малой дискриминантной силы алгоритма. Под «дискриминантной силой» понимается способность алгоритма разделять гипотезы. Достижение наилучшего разделения гипотез совсем не обязательно сопровождается увеличением правдоподобия правильной гипотезы. С другой стороны, максимизация $P(W_i|X,\lambda)$ в полном пространстве приводит к дискриминантному алгоритму, но требует вычисления всех или наиболее вероятных конкурирующих гипотез, представленных частью знаменателя $\sum_{k \neq i} P(X|W_k,\lambda)P(W_k)$, которую следует уменьшать.

Существуют различные техники дискриминантного (дискриминативного) обучения, максимизирующие $P(W_i|X,\lambda)$ – это «корректирующее обучение», «максимизация взаимной информации», «минимизация фонемной ошибки» или «минимизация ошибки классификации» [31, гл. 4]. Эти методы дают существенное улучшение распознавания на небольших словарях, однако при увеличении словаря дополнительные вычислительные затраты становятся чрезмерными, а преимущества незначительными.

В настоящее время получило распространение использование нейронных сетей, которые, помимо новых возможностей, как правило, создают модели, обладающие дискриминантными свойствами.

Рассмотрим современные системы, использующие «длительные» контекстные признаки и искусственную нейронную сеть в качестве классификатора.

10.1. Долговременные признаки

Долговременные признаки, или TRAP-признаки (TempoRAL Patterns) впервые были предложены в работах [39, 40]. В отличие от традиционных спектральных методов, основная идея предложенного метода заключается в использовании признаков, полученных путем объединения результатов вычисления энергии мел-фильтров на нескольких фреймах. Блок-диаграмма вычисления TRAP-признаков представлена на рис. 10.1. Процесс формирования долговременных TRAP-признаков начинается с вычисления энергии в мел-фильтрах. Речевой сигнал разбивается на фреймы длиной 25 мс с шагом 10 мс. Мел-фильтры эмулируются взвешиванием энергетического спектра речевого сигнала треугольными окнами, расположенными в логарифмическом масштабе. Значения, получаемые на выходе каждого из фильтров, затем суммируются и логарифмируются. Процедура повторяет получение MFCC коэффициентов (раздел 3.) за исключением кепстрального преобразования. Векторы TRAP-признаков формируются путем объединения нескольких значений внутри одной критической полосы мел-фильтра, то есть новый вектор признаков формируется из компоненты с одним номером нескольких последовательных векторов логарифма мел-спектра. Таким образом, TRAP-признаки – это набор независимых векторов, описывающих изменение во времени энергии речевого сигнала для каждой из критических полос мел-фильтра.

Дополнительно к данному вектору признаков можно применить нормализацию по среднему и СКО (среднеквадратическому отклонению), а также взвешивание с помощью окна Хэмминга. Полученный вектор признаков поступает затем на нейросетевой классификатор, выходы которого являются апостериорными вероятностями классов фонем или состояний фонем. Нейросетевой классификатор используется в каждой из критических полос мел-фильтра. Выходные данные нейросетевых классификаторов поступают на вход merger-сети – нейронной сети, целью которой является объединение результатов распознавания, независимо полученных на различных критических полосах мел-фильтра.

В описанных выше процедурах вычисляется вероятность появления фонемы для центрального фрейма вектора TRAP-признаков.

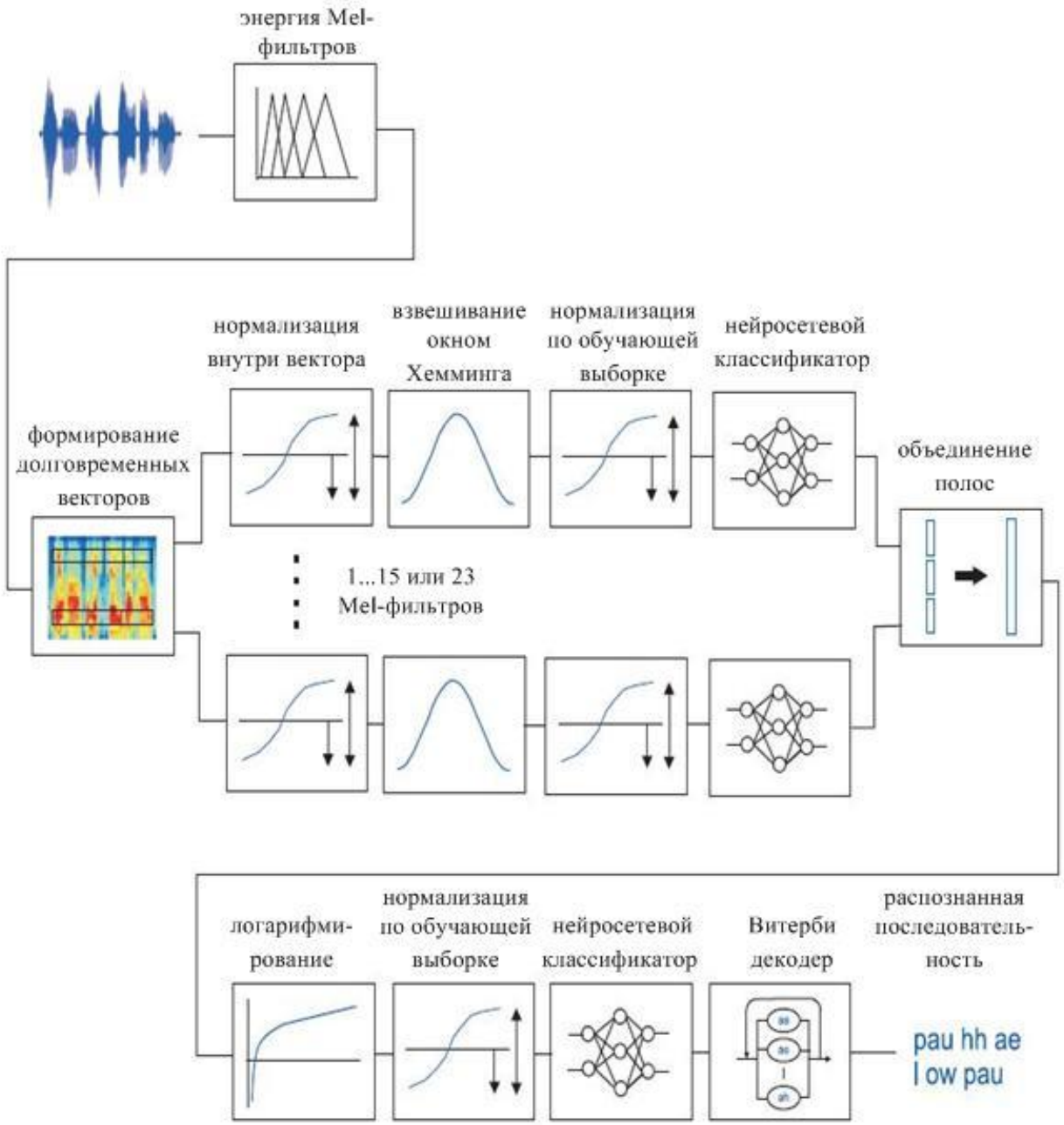


Рис. 10.1. Блок-диаграмма системы на основе TRAP-признаков [41].

Эксперименты с TRAP-признаками показали, что система, основанная на них, существенно превосходит стандартную систему, основанную на MFCC или PLP признаках – относительное уменьшение ошибки распознавания составило более 10%. Также было обнаружено, что ошибки систем с этими признаками не очень сильно коррелируют, что позволяет использовать их совместно, дополнительно уменьшая ошибку. С другой стороны, системы, основанные на данных признаках, требуют больших вычислительных затрат из-за использования нейросетевых классификаторов в каждой критической полосе, а также большого объема речевых баз данных для обучения.

Длина вектора признаков TRAP может достигать 101 элемента, что соответствует длительности контекста в 1 секунду. Эксперименты по определению оптимальной длины вектора на основе минимума фонемной

ошибки [41] дали длину в 300–400 мс, что соответствует размерности векторов TRAP-признаков в 31–41. (Обычно размерность выбирается нечётной, чтобы сопоставить центральному элементу вектора центр фонетического элемента и иметь одинаковые по длине контексты до и после центральной точки.) Принимая во внимание результаты этого эксперимента, следует учесть, что на оптимальную длину векторов TRAP-признаков может оказывать влияние объём обучающей выборки – выявление статистических закономерностей в пространстве большей размерности требует соответствующего увеличения базы данных. Очевидно, вариативность длительных временных признаков имеет ту же природу, что и вариативность трифонов (раздел 8.) – количество различных траекторий экспоненциально возрастает при увеличении контекстных зависимостей, при этом резко возрастает количество не наблюдённых в базе сочетаний фонем. Так, при использовании тестовой части базы данных TIMIT с 39 фонемами, при учёте биграмм, количество не встреченных сочетаний составляет 1104 или 2.26% от общего числа, при учёте триграмм, соответственно, 8952 и 18.83% и 4-грамм – 20681 и 54.55%. Этот эффект получил название «проклятье размерности» (the curse of dimensionality).

В настоящее время при исследовании 3–4-граммных сочетаний фонем используются базы данных, содержащие более 1000 часов речи. Отметим, что для русской речи не существует баз данных такого объема. Наиболее известные базы данных, такие как SpeechDat и SpeeCon, содержат порядка 60–70 часов.

Для преодоления этих трудностей было предложено [41] разбить контекст на левую и правую части. При этом части обучаются независимо друг от друга и объединяются потом на merger-сети. Также в [41] показано, что за счет разбиения долговременного контекста на две составляющие существенно снижается требование к объему обучающей выборки.

Структура системы с разделенным временным контекстом представлена на рис. 10.2. На первом шаге извлекается энергия из мел-фильтров и объединяется в 310 мс временной контекст (31 значение). Далее временной вектор из каждой критической полосы мел-фильтра разбивается на две части — левый контекст (значения 0–16) и правый контекст (значения 16–31). Обе части взвешиваются соответствующими половинами окна Хэмминга, после чего вычисляется дискретное косинусное преобразование (DCT). Для каждой части сохраняется *n* коэффициентов DCT-преобразования (число коэффициентов находится экспериментальным путем). Полученные для каждой критической полосы векторы объединяются в общий левый и правый контексты и подаются на два нейросетевых классификатора, которые формируют апостериорные вероятности аналогично TRAP-системам. Выходы нейросетевых классификаторов логарифмируются и совместно подаются на merger-сеть, которая также обучена формированию апостериорных вероятностей. В завершение апостериорные вероятности merger-сети декодируются с помощью Витерби-декодера, в результате чего получаются последовательности фонем. Данная система была названа авторами LC-RC-система (сокращение от Left context – Right context).

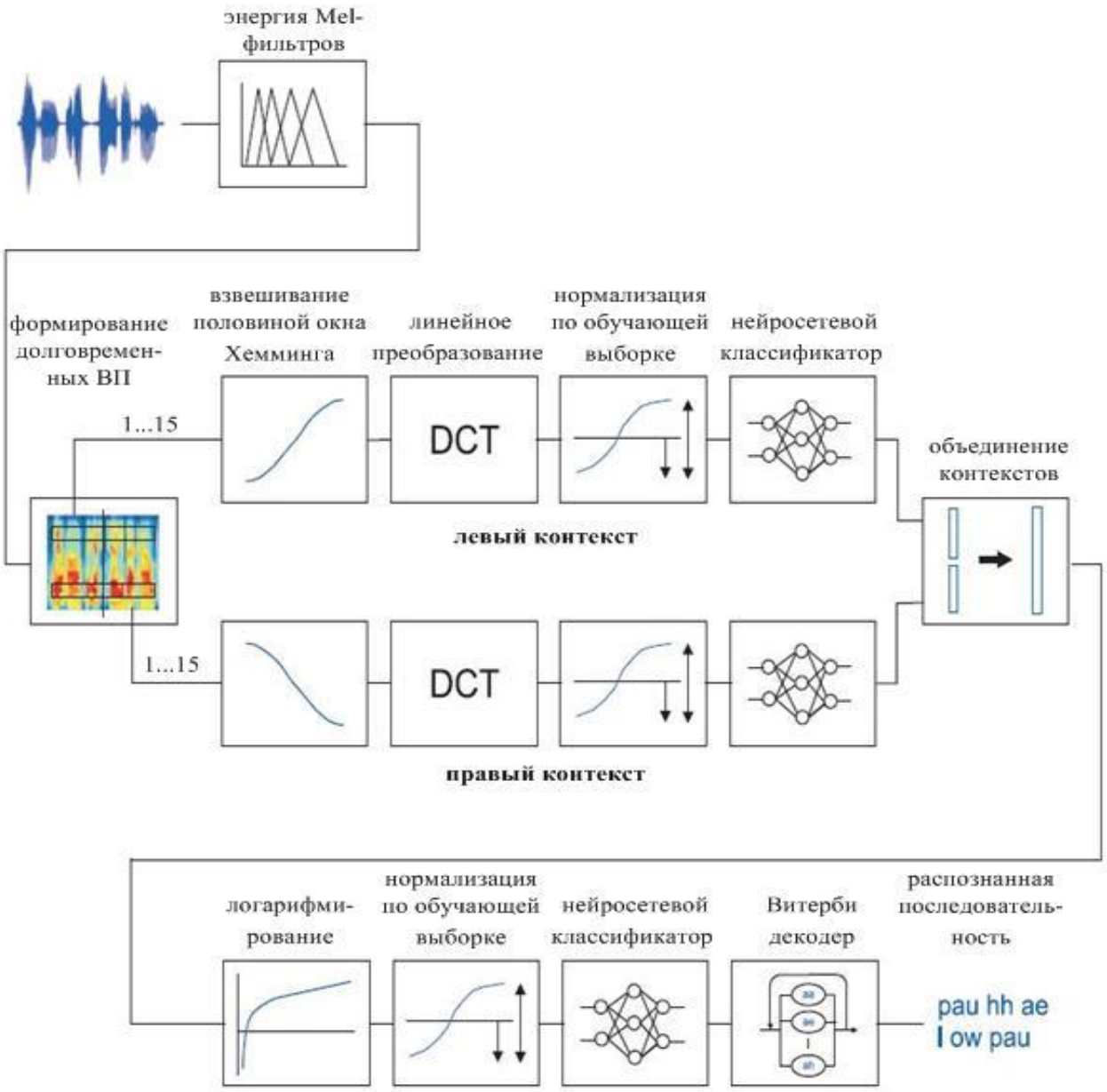


Рис. 10.2. Блок-диаграмма системы на основе LC-RC-признаков [41].

11. УСЛОВНЫЕ СЛУЧАЙНЫЕ ПОЛЯ

Условные случайные поля (CONDITIONAL RANDOM FIELDS, CRF) – это статистическая модель, иногда применяемая в настоящее время для распознавания образов. Первоначально модель возникла для объяснения ферромагнетизма [45]. Главным отличием этого подхода от метода скрытых марковских моделей (СММ) является то, что в рамках условных случайных полей не делается предположений о вероятностных распределениях моделируемых случайных процессов (как, впрочем, и в рамках моделей, использующих нейронные сети). Использование условных случайных полей сводится к задаче максимизации энтропии и вычислению условной вероятности речевого сегмента в виде экспоненциальной модели [46]:

$$p(w \mid x, \lambda) = \frac{1}{Z(x;\lambda)} \sum_{s \in w} \exp\{\bar{\lambda}^T \cdot \bar{f}(w, s, x)\}, \qquad (11.1)$$

где w – речевой фрагмент (слово, фраза…), s – состояния, x – набор векторов признаков, λ – вектор оптимизируемых при обучении параметров (множители Лагранжа), f – вектор достаточных статистик, $Z(x;\lambda) = \sum_{w,s \in w} \exp\{\bar{\lambda} \cdot \bar{f}(w, s, x)\}$ – нормирующий множитель, сумма по всем возможным речевым фрагментам.
Форма (13.1) вытекает из решения уравнения Лагранжа при самых общих предположениях относительно достаточных статистик.

В качестве достаточных статистик обычно фигурируют статистики первого и второго порядков:

$f_{ss'}^{(Tr)}(w, s, x) = \sum_{t=1}^{T} \delta(s_{t-1} = s)\delta(s_t = s')$ – статистика переходов между состояниями,

$f_s^{(Occ)}(w, s, x) = \sum_{t=1}^{T} \delta(s_t = s)$ – статистика состояний,

$f_{s,j}^{(M1)}(w, s, x) = \sum_{t=1}^{T} \delta(s_t = s) x_{j,t} \qquad j = 1,2,...N$ – статистика первого порядка,

$f_{s,j}^{(M2)}(w, s, x) = \sum_{t=1}^{T} \delta(s_t = s) x_{j,t}^2 \qquad j = 1,2,...N$ – статистика второго порядка, (11.2)

где δ – индикаторная функция (1 при выполнении условия, 0 – в противном случае), T – количество векторов признаков в речевом фрагменте, N – размерность пространства признаков.

Для того чтобы прояснить смысл вышеприведённых параметров, покажем, как будет выглядеть в этих терминах вероятность фрагмента в рамках стандартной модели СММ.

Полная вероятность речевого фрагмента в терминах СММ (раздел 6) имеет вид:

$$p(w \mid x) = \prod_{i=1}^{T} a_{s_{t-1},s_t} B(\bar{x}_t \mid s_t), \tag{11.3}$$

где s_t – состояние в момент времени t, $a_{ss'}$ – вероятность перехода из состояния s в состояние s', $B(x|s)$ – вероятность эмиссии (функция плотности вероятности) для состояния s.

Если функцию плотности вероятности аппроксимировать одной гауссовой функцией с диагональной матрицей ковариации, выражение (11.3) примет вид:

$$p(w \mid x) = \prod_{t=1}^{T} a_{s_{t-1},s_t} \frac{\exp(-\sum_{j=1}^{N} \frac{(x_{j,t} - \mu_j^{(s_t)})^2}{2\sigma_j^{(s_t)2}})}{(2\pi)^{\frac{N}{2}} \prod_{j=1}^{N} \sigma_j^{(s_t)}} =$$

$$\prod_{t=1}^{T} a_{s_{t-1},s_t} \frac{\exp(-\sum_{j=1}^{N} \frac{x_{j,t}^2 - 2x_{j,t}\mu_j^{(s_t)} + \mu_j^{(s_t)2}}{2\sigma_j^{(s_t)2}})}{(2\pi)^{\frac{N}{2}} \prod_{j=1}^{N} \sigma_j^{(s_t)}}. \tag{11.4}$$

где μ_j и σ_j – среднее и среднеквадратичное отклонения для компоненты с номером j гауссовой функции, аппроксимирующей функцию плотности вероятности для соответствующего состояния.

Выберем в качестве начального приближения для множителей Лагранжа следующие выражения [47]:

$$\lambda_{s,s'}^{(Tr)} = \log(a_{s,s'}),$$

$$\lambda_s^{(Occ)} = -\log(2\pi)^{\frac{N}{2}} \prod_{j=1}^{N} \sigma_j^{(s)} - \sum_{j=1}^{N} \frac{\mu_j^{(s)2}}{2\sigma_j^{(s)}},$$

$$\lambda_{s,j}^{(M1)} = \frac{\mu_j^{(s)}}{\sigma_j^{(s)}}, j = 1,...,N, \tag{11.5}$$

$$\lambda_{s,j}^{(M2)} = -\frac{1}{2\sigma_j^{(s)2}}, j = 1,...,N.$$

Можно проверить, что, если подставить выражения (11.5) и (11.2) в выражение (11.1), мы получим значение условной вероятности, тождественное (11.4) с точностью до нормирующего множителя. Таким образом, в рассмотренном случае итерации по уточнению множителей Лагранжа в процедуре обучения CRF начинаются с точки, в которой заканчиваются итерации алгоритмов обучения скрытой марковской модели, то есть результаты будут не хуже. Однако улучшение незначительно.

Использование моментов высокого порядка в рамках модели условных случайных полей очевидно. Можно также учесть модели языка, в этом случае, например, для униграммной модели достаточная статистика и начальное приближение множителей Лагранжа будет иметь вид:

$$\begin{aligned} f_{w'}^{(LM)}(w,s,x) &= \delta(w=w') \\ \lambda_{w'}^{(LM)} &= \log p_{w'} \end{aligned}, \qquad (11.6)$$

где p_w – вероятность появления слова w.

Достаточные статистики (11.2) приведены для случая аппроксимации функции плотности вероятности состояний одной гауссовой функцией. При этом предполагается, что в каждый момент времени система находится в одном состоянии в соответствии с алгоритмом Витерби. В случае, когда функция плотности вероятности аппроксимируется несколькими гауссовыми функциями и принадлежность к состоянию носит вероятностный характер в соответствии с алгоритмом Баума-Уэлша, достаточные статистики имеют следующий вид:

$$\begin{aligned} f_{s,m}^{(Occ)}(w,s,x) &= \sum_{t=1}^{T}\gamma_t(s_t=s,m),\ m=1,2,\ldots,M_s \\ f_{s,m,j}^{(M1)}(w,s,x) &= \sum_{t=1}^{T}\gamma_t(s_t=s,m)x_{j,t},\quad j=1,2,\ldots,N,\ m=1,2,\ldots,M_s \\ f_{s,m,j}^{(M2)}(w,s,x) &= \sum_{t=1}^{T}\gamma_t(s_t=s,m)x_{j,t}^2,\ j=1,2,\ldots,N,\ m=1,2,\ldots,M_s, \end{aligned} \qquad (11.7)$$

где M_s – количество гауссовых функций, применяющихся для аппроксимации функции плотности вероятности состояния s, $\gamma_t(s,m)$ – вероятность находиться в момент времени t в состоянии s и гауссовой функции m (6.25, 6.23).

12. НЕЙРОННЫЕ СЕТИ

Прежде чем перейти к рассмотрению ещё одной группы методов распознавания, остановимся на любопытном факте. Технологии распознавания речи являются достаточно молодыми, но очень перспективными в коммерческом смысле, что предполагает значительное финансирование и бурный рост. Однако, несмотря на хорошее финансирование, со времени, когда было предложено использовать марковские модели (середина шестидесятых годов XX века), прогресс в качестве распознавания на протяжении более 40 лет был довольно малым. Новым методам не удавалось преодолеть уровень результатов, достигнутых непрерывной марковской моделью с гауссовой аппроксимацией функций плотности вероятностей состояний, либо улучшение было столь незначительным, что не стоило существенного усложнения систем. При этом достигнутые результаты не позволяли использовать системы распознавания речи как массовый коммерческий продукт, хотя конкретные приложения в узких предметных областях уже давно работали.

Многим исследователям представлялось, что характер задачи соответствует возможностям искусственных нейронных сетей. Попытки использования нейронных сетей начались довольно давно. В качестве примера приведём статью 1990 г. [51], в которой было предложено много перспективных идей. В частности, использовались долговременные признаки (см. раздел 10.1) в виде одного супервектора, состоящего из 9 последовательных векторов мел-спектра, и рекуррентная связь между выходным и входным слоями, позволявшие учитывать контекстные зависимости. Несмотря на то, что в этой системе фактически использовались те же признаки, что и в стандартной марковской модели, плюс упомянутые усовершенствования, превзойти стандартную систему на основе гауссовых смесей не удалось. Этот факт вызвал такое недоумение в научной среде, что в 1996 вышла статья с красноречивым названием “Towards increasing speech recognition error rates” [52], в которой была сделана попытка объяснить длительное отсутствие прогресса в создании систем распознавания речи. Авторы объясняли отсутствие прогресса тем, что марковская модель на основе гауссовых смесей была принята в качестве базовой в десятках научных центров во всём мире и в течение нескольких лет была предельно оптимизирована, так что любой новой, сырой системе на начальном этапе превзойти её почти невозможно.

Несмотря на то, что приведённый аргумент трудно оспорить, последние работы, использующие многослойные нейронные сети различных типов, доказывают, что есть ещё одна элементарная причина – нейронные сети не обладали достаточной информационной мощностью, поскольку мощность компьютеров не позволяла использовать сети с несколькими слоями и выходным слоем, состоящим из нескольких тысяч нейронов, соответствующих трифонам (а не нескольким десяткам монофонов, как в ранних системах).

12.1. Глубокие нейронные сети

Нейронная сеть или перцептрон с любым количеством скрытых слоев является универсальным аппроксиматором [53], т.е. даже сети с одним скрытым слоем, использовавшиеся до этого этапа, могут аппроксимировать любую поверхность в пространстве признаков. Однако успех в распознавании речи пришел только с использованием многослойных сетей. Это объясняется невозможностью или крайней трудностью создания разумной методики инициализации весов для сетей с одним скрытым слоем, что приводит к далекому от оптимума набору весов при обучении.

Использование многослойных нейронных сетей поставило новую задачу – разработку новых алгоритмов обучения, поскольку известный алгоритм обратного распространения ошибки без разумной инициализации входных весов нейронов может приводить к неоптимальным решениям. По-видимому, разработка новых алгоритмов обучения будет трендом работ, связанных с использованием нейронных сетей.

Одним из методов является инициализация с помощью послойного обучения, начиная с нижних слоёв [54,55]. В качестве целевой функции для первого скрытого слоя рассматривается входной вектор признаков. Исходный вектор может содержать несколько последовательных MFCC или мел-спектральных векторов-признаков. Чтобы избежать тождественного преобразования, входной вектор зашумляют. Следующий слой нейронной сети таким же образом обучают воспроизводить выходные сигналы предыдущего слоя. Всего таким образом обучают до 5–7 слоёв. После того, как инициализация первых слоёв проведена, включают стандартный алгоритм обратного распространения ошибки для всей сети с целевой функцией, отражающей принадлежность входного сигнала к соответствующему трифону. Данный подход показал явное преимущество по сравнению с классическим подходом с гауссовыми смесями: результаты распознавания всегда оказывались лучше, причём многослойная сеть, обученная на речевом материале в 309 часов речи, показала лучшие результаты, чем метод с гауссовыми смесями, обученный на 2000 часах речи.

Любопытно отметить, что предлагаемый алгоритм обучения создаёт систему, напоминающую по функционированию слуховую. В слуховой системе обнаружены нейроны, реагирующие на определённые события в акустическом сигнале [56, гл. 9]. По мере «углубления» сигнала в центральные отделы слуховой системы характер признаков, выделяемых специализированными нейронами, принимает всё более сложный и избирательный характер. Предварительное обучение отдельных слоёв нейронной сети, выполняет ту же задачу – отдельные слои обучаются находить признаки сигнала всё более высокого уровня.

Если внутренние слои нейронных сетей выделяют признаки речевого сигнала, характерные для речи вообще, то их можно унифицировать для всех языков, обучая для каждого нового языка только выходной слой нейронной сети (рис. 12.1). Это было бы чрезвычайно важно, поскольку для обучения только

одного слоя нейронной сети требовалась бы гораздо меньшая речевая база данных, чем для обучения всех 5–7 слоев.

Эксперименты полностью подтвердили такую возможность. Использование совместно речевых баз данных для французского, немецкого и итальянского языков позволило уменьшить ошибку распознавания на 3,3–5,4% в относительном выражении по сравнению с моноязыковыми моделями [55].

Следует отметить, что содержащаяся во внутренних слоях нейросетей информация о признаках речевого сигнала может быть использована для распознавания речи на неродственных языках. Были поставлены эксперименты по использованию внутренних слоев нейронной сети, обученной на базе данных европейских языков, для дообучения выходного слоя нейросети для китайского языка. Относительный выигрыш составил от 21,1% до 8,3% при увеличении базы данных китайского языка от 3 до 139 часов [55]. Рассмотренный прием открывает возможность создавать системы распознавания для всевозможных языков, в том числе малоресурсных.

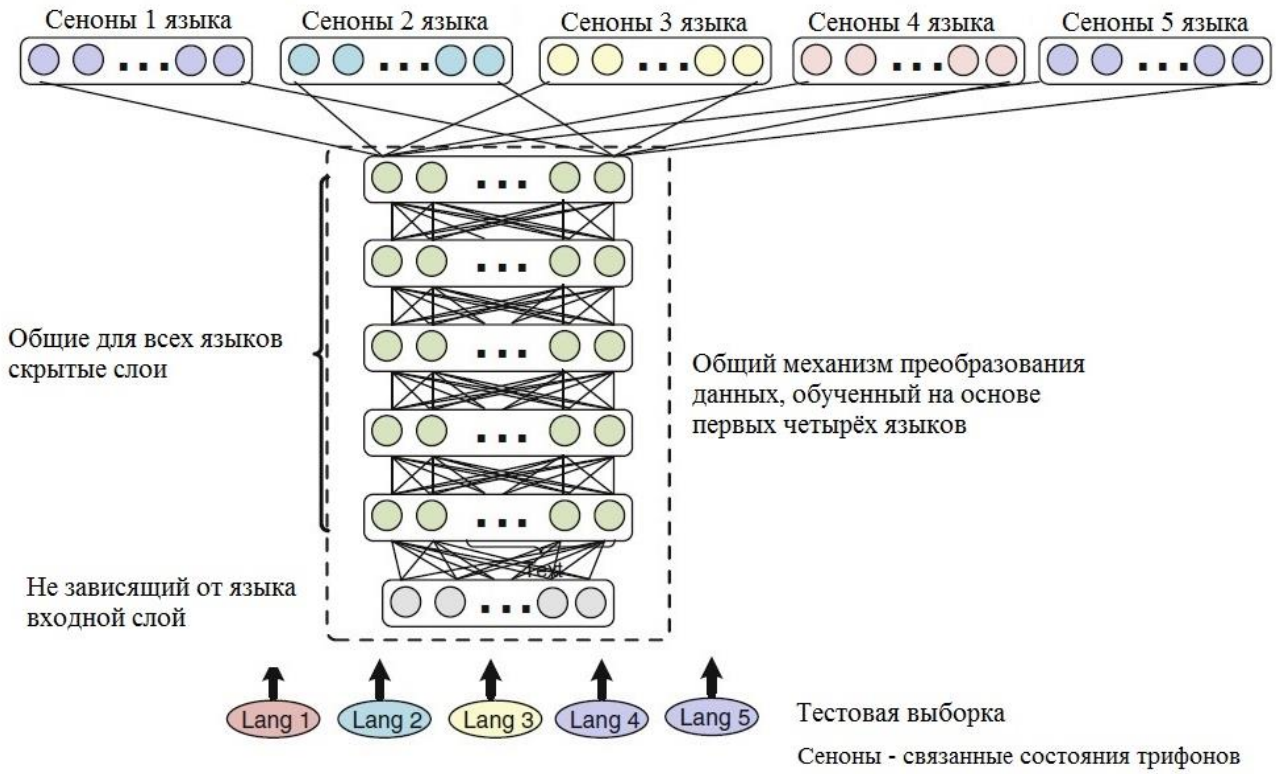


Рис. 12.1. Обучение системы, обученной четырем языкам, пятому языку [55].

Поскольку нейронные сети не могут идентифицировать динамические объекты, для сравнения моделей с сигналом по-прежнему используется формализм марковских моделей, однако теперь в качестве вектора признаков используется набор апостериорных вероятностей трифонов, полученный на выходе нейронной сети. Такой метод использования нейронных сетей одним из первых предложил для монофонов Х. Германский с соавторами [57].

Отметим, что контекстная зависимость, то есть влияние фонем друг на друга, в данном случае моделируется построением входного вектора из нескольких последовательных векторов признаков, описывающих отрезок сигнала длиной

около 25 мс. Окна анализа смещаются на 10 мс. Таким образом, для того чтобы отобразить отрезок сигнала длиной 300 мс (такие размеры контекстной зависимости были выявлены в работе [41]), требуется около 30 векторов признаков. Таким образом, размерность результирующего супервектора может составлять от 300 до более чем 1000, в зависимости от размерности исходного вектора признаков. Работа с векторами такой размерности требует большого количества вычислений.

Более существенным недостатком, присущим данному методу, является то, что глубокие нейронные сети не могут распознавать динамические объекты, из-за чего и приходится использовать алгоритм Витерби в рамках марковской модели. Недостатки марковской модели довольно очевидны: дискретность, или независимость последовательных состояний друг от друга; отсутствие глубоких временных связей, то есть неспособность распознавать траектории в пространстве признаков как информативные объекты.

12.2. Рекуррентные нейронные сети

Можно предположить, что оба отмеченных недостатка можно преодолеть, используя рекуррентные нейронные сети. Рекуррентные нейронные сети содержат нейроны, объединенные в направленный круговой процесс. Это наделяет нейронную сеть памятью и, следовательно, способностью распознавать процессы, а не только статические объекты, как рассмотренные выше глубокие нейронные сети. Рекуррентные нейронные сети отличаются от рассмотренных ранее многослойных тем, что при обработке очередного вектора признаков система учитывает также внутренние состояния нейронов, которые, в свою очередь, формируются предыдущими векторами признаков и состояниями в предыдущие моменты времени. В этом смысле единичная рекуррентная нейронная сеть представляет собой более мощное образование, чем глубокая нейронная сеть. Тем не менее, рассматриваются иерархические комбинации рекуррентных нейронных сетей и комбинации рекуррентных и многослойных сетей. Это можно объяснить, как и для многослойных сетей, желанием структурировать систему, приблизить её к принципам функционирования нервной системы, упростить процедуру инициализации и обучения.

Надо сказать, что совместное использование глубоких и рекуррентных нейронных сетей открывает много возможностей для создания различных архитектур систем распознавания. Выбор «правильной» архитектуры может существенно улучшить качество распознавания. Так, в работе [58] рекуррентная нейронная сеть располагается после «языкового» слоя многослойной сети, или между выходным, «языковым», слоем и последним внутренним слоем (рис.12.1). В качестве признаков используются апостериорные вероятности фонем с «языкового» слоя и (или) выходы последнего слоя глубокой нейронной сети. Такую архитектуру можно считать грубой моделью восходящего слухового пути и нижнего уровня центральной слуховой системы. Хотя качество распознавания по сравнению с рассмотренной выше глубокой нейронной сетью (без рекуррентной) увеличилось всего на 0,4%-0,5% (абсолютно), но сама

рекуррентная сеть улучшила точность распознавания фонем с 71,8% до 81,2% по сравнению с архитектурой, где многослойная сеть отсутствовала и на вход рекуррентной подавались исходные спектральные признаки.

Наряду с поисками архитектуры системы распознавания большое внимание уделяется усовершенствованию рекуррентных сетей. Дело в том, что, несмотря на улучшение качества распознавания и возможность обходиться без искусственного метода скрытых марковских моделей, они обладают и недостатками. Главным недостатком является трудоёмкость и сложность процедуры обучения. Обычно применяется всё тот же метод обратного распространения ошибки, но, учитывая рекуррентность, развёрнутый во времени - back-propagation-through-time (BPTT). Основные трудности при этом связаны с обнулением градиента и очень большим временем обучения [59,60].

Рассмотрим основные усовершенствования рекуррентных нейронных сетей, использующиеся в системах распознавания речи.

Обычная рекуррентная нейронная сеть описывается уравнениями:

$$\mathbf{h}_t = f(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}) \quad (12.1)$$

$$\mathbf{y}_t = g(\mathbf{W}_{hy}\mathbf{h}_t) \quad (12.2)$$

где: $\mathbf{x}_t$, $\mathbf{h}_t$ и $\mathbf{y}_t$ – вектор признаков, вектор внутренних состояний нейронов и выходной вектор, соответственно, в момент времени t; $\mathbf{W}_{xh}$ – матрица весовых коэффициентов, связывающая входной вектор с вектором внутренних состояний; $\mathbf{W}_{hh}$ – матрица, связывающая векторы внутренних состояний; $\mathbf{W}_{hy}$ – матрица, связывающая вектор внутренних состояний с выходным вектором; f – сигмоидная функция; g – линейная или softmax функция (напомним, что softmax функция гарантирует нормированность выходного вектора на 1, что позволяет рассматривать компоненты выходного вектора как вероятности). Размерность вектора состояний $\mathbf{h}$ равна числу нейронов в сети. Размерности матриц очевидны, исходя из размерностей векторов, которые они связывают.

В рекурсии (12.1) может участвовать также выходной вектор, тогда формула (12.1) приобретает вид:

$$\mathbf{h}_t = f(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{yh}\mathbf{y}_{t-1}), \quad (12.3)$$

где $\mathbf{W}_{yh}$ – матрица, связывающая выходной вектор с вектором внутренних состояний.

Обучение такой нейронной сети, как обычно, заключается в оптимизации всех или части матриц весовых коэффициентов с целью достижения оптимума по некоторому критерию, зависящему от выходного вектора и целевого вектора для данного момента времени по всей обучающей выборке.

Одним из самых популярных вариантов рекуррентных нейронных сетей, использующихся для распознавания речи, являются сети с нейронами с «длинной кратковременной памятью» (Long Short Term Memory, LSTM) [61-66]. Нейроны в такой сети имеют более сложную структуру и содержат блок памяти, включающий гейты (gates) входа, забывания и выхода со своими векторами состояний. Такая нейронная сеть описывается следующим набором уравнений:

$$\mathbf{i}_t = f(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i) \quad (12.4)$$

$$\mathbf{f}_t = f(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f) \quad (12.5)$$

$$\mathbf{c}_t = \mathbf{f}_t \bullet \mathbf{c}_{t-1} + \mathbf{i}_t \bullet \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_{c)} \quad (12.6)$$

$$\mathbf{o}_t = f(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t + \mathbf{b}_o) \quad (12.7)$$

$$\mathbf{h}_t = \mathbf{o}_t \bullet \tanh(\mathbf{c}_t) \quad (12.8)$$

$$\mathbf{m}_t = \mathbf{o}_t \bullet \mathbf{h}_t \quad (12.9)$$

$$\mathbf{y}_t = \sigma(\mathbf{W}_{ym}\mathbf{m}_t + \mathbf{b}_y), \quad (12.10)$$

где добавляются векторы состояний $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{c}_t$ и $\mathbf{o}_t$ – входа, забывания, активации и выхода, соответственно, $\mathbf{m}_t$ – вспомогательный вектор, ● – поэлементное произведение векторов, **b** – векторы смещений, σ – логистическая сигмоидальная функция, смысл матриц весовых коэффициентов очевиден, все они полные, кроме $\mathbf{W}_{ci}$, которая диагональна. Размерность всех внутренних векторов одинакова и равна числу нейронов в сети.

Проблемы с экспоненциальным уменьшением градиента при обучении в данной нейронной сети обходятся благодаря тому, что сигнал ошибки захватывается блоком памяти и продолжает воздействовать на другие гейты, даже если текущий сигнал ошибки исчез.

Ещё один тип рекуррентных нейронных сетей основан на резервуарной модели [67-70]. Это простейший вариант рекуррентной сети, основанный на нейронах с утечкой, описывающийся уравнениями, аналогичными (12.1 и 12.2):

$$\mathbf{h}_t = (1-\lambda)\mathbf{h}_{t-1} + \lambda f(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}) \quad (12.11)$$

$$\mathbf{y}_t = \mathbf{W}_{hy}\mathbf{h}_t, \quad (12.12)$$

где коэффициент λ определяет скорость утечки.

Выходной вектор является линейной комбинацией компонент вектора состояний. Все компоненты матрицы $\mathbf{W}_{hh}$ инициализируются случайными числами и не меняются при обучении. Целью обучения является оптимизация матрицы $\mathbf{W}_{hy}$. Благодаря линейности выхода решение возможно в аналитическом виде, без привлечения алгоритма BPTT:

$$\mathbf{W}_{hy} = (\mathbf{X}^T\mathbf{X} + \varepsilon\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{D}), \quad (12.13)$$

где **X** – матрица, строки которой образованы из векторов $\mathbf{h}_t$ для обучающей выборки, **D** – матрица, строки которой образованы из соответствующих целевых векторов, ε – константа регуляризации, **I** – единичная матрица.

Вместо формулы (12.11) можно использовать обычную формулу softmax (12.2). В этом случае аналитическое решение невозможно, и привлекается метод градиентного спуска. Такой подход имеет смысл, когда количество нейронов в сети столь велико, что обращение матрицы в формуле (12.13) становится затруднительным.

Несмотря на крайнюю простоту таких сетей, результаты распознавания, полученные с их помощью, не сильно уступают результатам сетей LSTM.

Отметим также такую разновидность рекуррентных сетей, как autoregressive moving average (ARMA) рекуррентные сети. В этих сетях в качестве входного вектора используется супервектор, состоящий из нескольких векторов в последовательные моменты времени. Понятно, что преимущество низкой размерности входного вектора при этом теряется. В работе [58] используются 13 последовательных векторов. Процент правильно распознанных фонем по

сравнению с канонической рекуррентной сетью увеличился с 80.2% до 81.2%. Можно предположить, что за счёт оптимизации общей архитектуры системы распознавания можно добиться большего прогресса.

12.3. Нормализация и адаптация нейронных сетей

Проведение параллели между системами распознавания, основанными на нейронных сетях и слуховой системой человека, может привести к неправильному умозаключению, что эти системы не нуждаются в адаптации к диктору, каналу или шумовому окружению, поскольку человек явным образом не адаптируется к новым условиям. Однако это не так, по крайней мере по двум причинам. Во-первых, мощность нейронных сетей, используемых для распознавания, пока ещё очень сильно уступает мощности естественных нейронных сетей. Во-вторых, объём речевого материала, получаемый ребёнком за 2-4 года, пока его восприятие речи не приблизится по качеству к восприятию взрослого, гораздо больше, чем объём обучающей базы данных, используемый в системах распознавания.

Возможно, существует ещё одна причина. Дело в том, что овладение фонетической [93] и мелодической [94] структурами языка осуществляется в самом раннем детстве и ещё в утробе матери не путём обучения, а путём «импринтинга» [95]. Под импринтингом понимается заполнение отфильтрованными признаками речевого сигнала готовых нейронных структур, полученных путём наследования. В отличие от обучения, которое требует повторения и сознательных усилий, импринтинг осуществляется на инстинктивном уровне. Далее, в течение жизни, «без употребления полученные знания быстро утрачиваются» [95], а полученные импринтингом сохраняются, даже если ребёнок в очень раннем детстве был усыновлён и получил другой язык в качестве родного [93]. Впрочем, новый язык также усваивается импринтингом, если переселение произошло в достаточно раннем возрасте.

На основании вышеизложенного можно задать вопрос: любую ли задачу распознавания можно решить с помощью произвольной нейронной сети, или некоторые задачи требуют специальной архитектуры? Известно, что для распознавания статических образов достаточно нейронной сети с одним скрытым слоем (для распознавания линейно разделимых образов достаточно перцептрона Розенблатта с линейной активационной функцией); для распознавания динамических объектов с произвольным темпом уже имеет смысл использовать рекуррентную сеть (раздел 12.2). Возможно, «предустановленные» нейронные сети, заполненные с помощью импринтинга, умеют выделять в каком-то смысле инвариантные признаки речевого сигнала, для чего произвольным нейронным сетям может потребоваться гораздо большее число нейронов, превосходящее возможности современных компьютеров.

Таким образом, задача адаптации для распознавания речи на основе нейронных сетей остаётся актуальной.

Адаптация моделей на основе гауссовых смесей (GMM) заключается в смещении и иногда изменении формы функций плотности вероятности фонем в

пространстве признаков (раздел 9), как правило, на основе критерия максимума правдоподобия. Модели на основе нейронных сетей являются дискриминантными (а не генеративными, как GMM), то есть для их построения используется критерий качества распознавания (раздел 10), поэтому они не содержат структур наподобие функций плотности вероятности, которые смещаются целиком при изменении источника речи – информация в них распределена в тысячах весовых коэффициентов неявным образом. Поэтому рассмотренные в разделе 9 методы адаптации не могут быть использованы напрямую.

Что же касается метода нормализации признаков по длине голосового тракта (раздел 9.6) и других методов нормализации, то ничто не мешает его использованию. Напомним, что выбор искажающей частотной шкалы в данном методе опирается на результаты распознавания с помощью нескольких вариантов признаков, полученных с различными шкалами.

Прежде чем рассмотреть методы адаптации, применимые к системам распознавания речи на основе произвольных нейронных сетей, рассмотрим предложенный в [96] способ, позволяющий использовать для нейронных сетей все рассмотренные в разделе 9 методы адаптации, но ценой некоторых ограничений.

Существующие системы распознавания, основанные на нейронных сетях, используют в качестве входных признаков MFCC коэффициенты или мел-спектр (раздел 3). В работе [96] в качестве признаков предлагается использовать, после некоторых преобразований, вектор, компонентами которого являются вероятности всех состояний монофонов, вычисленные с помощью GMM моделей (раздел 6). Модели GMM обучаются с помощью скрытых марковских моделей (HMM) обычным образом. Далее размерность вектора из вероятностей монофонов уменьшается с помощью метода главных компонент и расширяется путём конкатенации 11 последовательных векторов до размерности 550, то есть описывает сигнал во временном окне 110 мс. Этот вектор и является входным для глубокой нейронной сети. Роль скрытых марковских моделей в данной схеме вспомогательная – после того, как функции плотности вероятности состояний, аппроксимируемые гауссовыми смесями, построены, они больше не используются. Однако этот механизм позволяет применить весь арсенал методов адаптации, наработанный для стандартной схемы HMM-GMM (раздел 9).

Попробуем оценить этот метод качественно. Успех нейронных сетей по сравнению с методом HMM-GMM можно объяснить тем, что нейронные сети находят оптимальную процедуру распознавания исходя из акустических закономерностей, выделяемых из большой речевой базы данных, и используя только критерий качества распознавания, а метод HMM-GMM основан на представлениях о речи, полученных косвенным путём. Например, при получении MFCC признаков, шкала мел (раздел 2.2) основана на психоакустических экспериментах, логарифмирование энергии в полосах опирается на представления об обработке сигналов нейронами (рис. 2.11), а косинусное преобразование не имеет физиологического обоснования вообще. Далее распознавание опирается на концепцию функций плотности вероятности в не

вполне адекватном признаковом пространстве. Хотя признаки MFCC используются и в нейронных сетях, суть сказанного заключается в том, что при использовании нейронных сетей имеет смысл избавляться от любых построений, не оптимизируемых прямо с помощью критерия качества распознавания. Так, нейронные сети находят способ распорядиться логарифмом спектра в мел-полосах лучше, чем алгоритм MFCC, что было показано в работах [55,97]. Можно предположить, что нейронные сети также смогли бы найти лучшие с точки зрения распознавания речи границы спектра, чем шкала мел, или нашли способ обходиться вообще без фиксированных границ. Можно ожидать, что на мощную нейронную сеть в будущем можно будет подавать непосредственно огибающую речевого сигнала с минимальной предварительной обработкой, то есть нейронные сети смогут реализовать нелинейное спектральное преобразование. В рассмотренном методе адаптации используются не только признаки MFCC, но и метод распознавания HMM-GMM, то есть нейронной сети предлагается распознавать речь на основе данных, полученных методом «вчерашнего дня», в котором, возможно, уже потеряна некоторая часть дискриминантной информации.

На основе сказанного можно признать, что данный метод адаптации является очень остроумным, но временным решением проблемы, актуальным при современном уровне мощности нейронных сетей.

Рассмотрим несколько методов адаптации к диктору, применяемых непосредственно к нейронным сетям. Адаптация к шумовому окружению и каналу связи осуществляется сходным образом. Будем придерживаться классификации, предложенной в монографии [55] – методы линейного преобразования, методы ограниченного обучения и методы подпространств.

12.3.1. Методы линейного преобразования

Данные методы объединяет то, что в уже обученную дикторонезависимую глубокую нейронную сеть добавляют ещё один слой, в котором происходит линейное преобразование признаков, то есть нейроны слоя имеют линейную, а не сигмоидальную, функцию активации (рис. 2.11). Весовые коэффициенты линейного слоя инициализируются единичными матрицами, а смещение - нулём. Используется обычный для глубоких нейронных сетей критерий обучения – минимизации ошибок. При этом обучению подвергаются только нейроны линейного слоя.

Линейный слой можно внедрить перед первым скрытым слоем нейронной сети, при этом линейной трансформации подвергаются входные признаки. Такой вариант нейронной сети называется LIN (Linear Input Network), или fDLR (Feature Discriminative Linear Regression). Если линейный слой внедряется перед выходным (softmax) слоем, сеть называют LON (Linear Output Network). Если линейный слой внедряется между скрытыми слоями, сеть называют LHN (Linear Hidden Network).

Для LON и LHN возможны два варианта подключения линейного слоя: до применения весовых коэффициентов исходной нейронной сети (12.14), или после них (12.15).

$$\mathbf{s}_{lin}^{n} = \mathbf{W}^{n}\mathbf{x}_{lin}^{n-1} + \mathbf{b}^{n} = \mathbf{W}^{n}(\mathbf{W}_{lin}\mathbf{x}^{n-1} + \mathbf{b}^{lin}) + \mathbf{b}^{n} = (\mathbf{W}^{n}\mathbf{W}_{lin})\mathbf{x}^{n-1} + (\mathbf{W}^{n}\mathbf{b}^{lin} + \mathbf{b}^{n}) \quad (12.14)$$

$$\mathbf{s}_{lin}^{n} = \mathbf{W}_{lin}\mathbf{s}^{n} + \mathbf{b}^{lin} = \mathbf{W}_{lin}(\mathbf{W}^{n}\mathbf{x}^{n-1} + \mathbf{b}^{n}) + \mathbf{b}^{lin} = (\mathbf{W}_{lin}\mathbf{W}^{n})\mathbf{x}^{n-1} + (\mathbf{W}_{lin}\mathbf{b}^{n} + \mathbf{b}^{lin}), \quad (12.15)$$

где n-1 и n – номера слоёв исходной нейронной сети, между которыми внедряется линейный слой, $\mathbf{W}^{n}$ – матрица весовых коэффициентов исходной нейронной сети слоя n, $\mathbf{W}_{lin}$ - матрица весовых коэффициентов линейного слоя, $\mathbf{b}^{n}$ и $\mathbf{b}^{lin}$ – соответствующие векторы смещений, $\mathbf{x}^{n-1}$ – вектор признаков, полученный на выходе слоя n-1, $\mathbf{s}^{n}$ и $\mathbf{s}^{n}_{lin}$ – соответствующие векторы состояний после суммирования (рис.2.11).

Размер матрицы $\mathbf{W}_{lin}$ для этих типов подключения может существенно отличаться. Допустим, количество нейронов в слое n-1 равно N_{n-1} а в слое n - N_n, тогда нетрудно увидеть, что в обоих случаях матрицы $\mathbf{W}_{lin}$ квадратные, но в первом случае (12.14) размерность матрицы равна N_{n-1}, а во втором (12.15) - N_n. Если размер слоя n-1 намного меньше, чем слоя n, как бывает, когда используют слой «бутылочное горло», то в первом случае обучать придётся гораздо меньше параметров, чем во втором. Аналогичные рассуждения в пользу второго случая справедливы, если линейный слой разместить перед бутылочным горлом.

Выбор подключения зависит от величины речевой базы для адаптации – как уже говорилось выше, лучше надёжно обучить небольшое количество параметров, чем пытаться обучать большое число параметров, не имея достаточного материала для дообучения.

Для LIN сети также имеется возможность резко уменьшить количество обучаемых параметров в матрице $\mathbf{W}_{lin}$, если дополнительная речевая база мала. Напомним, что входной вектор признаков представляет собой супервектор большой размерности, полученный конкатенацией нечётного числа векторов признаков (фреймов), полученных на окнах длительностью 16-20 мс. Линейный слой можно применять к отдельным фреймам этого супервектора. Нетрудно подсчитать, что, если количество фреймов равно 2n+1, то матрица коэффициентов линейного слоя уменьшится в $(2n+1)^2$ раз по сравнению с «полным» вариантом.

В результате адаптации к каждому новому диктору будет создаваться матрица $\mathbf{W}_{lin}$ и вектор смещения $\mathbf{b}^{lin}$ данного диктора, которые нужно хранить, если мы хотим в будущем подключать уже принятого в систему диктора без адаптации. Предпочтительнее было бы создать некоторый набор этих параметров для быстрого подключения любого диктора к системе. Такой метод реализован в адаптации методом подпространств (см. ниже).

12.3.2. Методы ограниченного обучения

Методы линейного преобразования обладают не очень хорошей способностью к адаптации именно в силу своей линейности. Их очевидное достоинство – простота и скорость. Однако, если необходимо добиться более

качественного результата, и адаптационная база речи это позволяет, имеет смысл пытаться модифицировать всю нейронную сеть. Стандартный способ обучения здесь неприменим, поскольку адаптационная база речи обычно слишком мала для того, чтобы настроить гигантское количество параметров. Следовательно, в процедуру обучения необходимо включить какие-то механизмы, ограничивающие возможность «испортить» уже обученную дикторонезависимую систему. Самое простое – при обучении принимать изменения не всех коэффициентов нейронной сети, а только тех, которые изменились в результате обучения наиболее сильно. Ранжируя коэффициенты по величине изменения, можно подобрать их оптимальное количество. Можно также уменьшить скорость обучения и количество циклов обучения. Другим способом является регуляризация с помощью добавки, связанной с различием новых и старых параметров нейронной сети, к адаптационному критерию. Адаптация будет происходить до тех пор, пока улучшение критерия превосходит изменение параметров. Рассмотрим два таких способа.

L2 регуляризация.

Составим супервектор из всех параметров нейронной сети – столбцов всех матриц весовых коэффициентов и векторов смещений. Пусть $\mathbf{W}_{SI}$ - такой супервектор для исходной дикторонезависимой нейронной сети, а $\mathbf{W}$ – супервектор для адаптированной сети. Введём L_2 норму:

$$R_2(\mathbf{W}_{SI} - \mathbf{W}) = \left\| vec(\mathbf{W}_{SI} - \mathbf{W}) \right\|_2^2 = \sqrt{\sum_{i=1}^{L} (W_{SI}^i - W^i)^2}\ , \qquad (12.16)$$

где L – размерность супервекторов, i – номер компоненты вектора.

Очевидно, что R_2 растёт по мере увеличения различий между исходной и адаптированной сети.

К обычному оптимизируемому критерию качества распознавания, например, количеству ошибок, добавим норму (12.16):

$$J_{L2}(\mathbf{W},\mathbf{b};S) = J(\mathbf{W},\mathbf{b};S) + \lambda R_2(\mathbf{W}_{SI},\mathbf{b}_{SI};\mathbf{W},\mathbf{b}), \qquad (12.17)$$

где $\mathbf{W}$ и $\mathbf{b}$ – параметры адаптированной нейронной сети, $\mathbf{W}_{SI}$ и $\mathbf{b}_{SI}$ параметры исходной дикторонезависимой сети, $S=\{(\mathbf{x}^m,\mathbf{y}^m)|0\le m<M\}$ – адаптационная речевая база ($\mathbf{x}^m$ – набор входных векторов, $\mathbf{y}^m$ – соответствующий набор целевых векторов вероятностей состояний, M - количество векторов), λ – оптимизируемый параметр.

Обучение на адаптационной речевой базе продолжается до тех пор, пока видоизменённый критерий уменьшается.

Регуляризация Кульбака-Лейблера.

В данном методе критерий качества распознавания дополняется расстоянием Кульбака-Лейблера между распределениями вероятности состояний для дикторонезависимой и адаптированной сетей:

$$J_{KLD}(\mathbf{W},\mathbf{b};S) = (1-\lambda)J(\mathbf{W},\mathbf{b};S) + \lambda R_{KLD}(\mathbf{W}_{SI},\mathbf{b}_{SI};\mathbf{W},\mathbf{b};S), \qquad (12.18)$$

где R_{KLD} - расстояние Кульбака-Лейблера:

$$R_{KLD}(\mathbf{W}_{SI},\mathbf{b}_{SI};\mathbf{W},\mathbf{b};S)=\frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{C}P_{SI}(i\mid\mathbf{x}_k;\mathbf{W}_{SI},\mathbf{b}_{SI})\log P(i\mid\mathbf{x}_k;\mathbf{W},\mathbf{b}), \qquad (12.19)$$

где $\mathrm{P_{SI}(i|\mathbf{x}_k;\mathbf{W}_{SI},\mathbf{b}_{SI})}$ и $\mathrm{P(i|\mathbf{x}_k;\mathbf{W},\mathbf{b})}$ – вероятности наблюдения m принадлежать к состоянию i, полученные дикторонезависимой и адаптированной нейронной сетью, соответственно.

Если используется критерий кросс-энтропии, то:

$$J(\mathbf{W},\mathbf{b};\mathbf{x},\mathbf{y})=-\frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{C}P_{emp}(i\mid\mathbf{x}_k)\log(P(i\mid\mathbf{x}_k)), \qquad (12.20)$$

где $P_{\mathrm{emp}}(\mathrm{i}|\mathbf{x}_\mathrm{k})$ – наблюдаемая на адаптационной выборке вероятность того, что наблюдение $\mathbf{x}_\mathrm{k}$ принадлежит состоянию i.

С использованием данного критерия, дополненного расстоянием Кульбака-Лейблера, процесс адаптации завершится, когда уменьшение кросс-энтропии в результате дообучения сравняется с увеличением расстояния между распределениями вероятностей состояний дикторонезависимой и адаптированной нейронных сетей.

Методы ограниченного обучения ставят проблему сохранения параметров системы ещё более остро, чем линейные методы – в рассмотренном виде надо хранить все параметры нейронной сети для каждого диктора. Конечно, можно адаптировать только один слой нейронной сети, но эксперименты показали, что адаптация в таком случае может значительно уступать полному варианту.

Поскольку матрицы коэффициентов адаптированной к диктору нейронной сети не очень сильно отличаются от матриц коэффициентов дикторонезависимой сети, можно ожидать, что дельта матрицы (разности соответствующих матриц) имеют низкий ранг и могут быть аппроксимированы матрицами низкой размерности стандартными методами с помощью сингулярного разложения [98]. Если ограничиться небольшим количеством наибольших сингулярных коэффициентов, порядок матриц можно существенно уменьшить. Таким образом можно уменьшить объём сохраняемой для каждого диктора информации в 10 раз без существенного ухудшения качества адаптации.

Ещё один метод предполагает, что исходные весовые матрицы имеют низкие ранги. Тогда каждую такую матрицу можно факторизовать:

$$\mathbf{W}_{m\times n}=\mathbf{W}^{2}_{m\times r}\mathbf{W}^{1}_{r\times n}, \qquad (12.21)$$

где $\mathbf{W}_{\mathrm{m\times n}}$ – исходная матрица размерности m×n и r<<m, r<< n.

Тогда можно внедрить между матрицами $\mathbf{W}^2$ и $\mathbf{W}^1$ матрицу $\mathbf{W}^3{}_{\mathrm{r\times r}}$, как это делалось в линейном методе, и адаптировать только её. Для исходной дикторонезависимой сети матрицы $\mathbf{W}^3{}_{\mathrm{r\times r}}$ единичны. Таким методом можно уменьшить объём запоминаемой информации для каждого диктора в 100 раз.

Рассмотренная процедура адаптации естественным образом приводит к ещё одному методу, который уже не связан с рангом исходной матрицы. Применим

сингулярное разложение к матрице коэффициентов между некоторыми слоями сети:

$$\mathbf{W}_{m\times n} = \mathbf{U}_{m\times m}\mathbf{\Sigma}_{m\times n}\mathbf{V}^{T}_{n\times n}\ , \qquad (12.22)$$

где **U** и **V** – унитарные матрицы, состоящие из левых и правых сингулярных векторов, **Σ** – матрица, у которой на главной диагонали находятся сингулярные числа, а все остальные элементы равны нулю.

Будем подвергать адаптации только диагональ матрицы **Σ**, более того, если речевого материала для адаптации недостаточно, будем адаптировать несколько верхних сингулярных чисел (количество определяем эмпирически). Поскольку запоминать требуется только диагональные элементы, а количество нейронов в слоях сети может достигать нескольких сотен, экономия становится ещё существеннее.

12.3.3. Методы подпространств

В результате использования любых рассмотренных выше методов адаптации мы получаем некоторое количество матриц и векторов, которые позволяют улучшить распознавание речи для одного диктора. Если мы адаптируем таким образом дикторонезависимую систему распознавания для S дикторов, и S достаточно велико, можно пытаться построить пространство дикторов, в котором каждый диктор будет представлен точкой. Для этого из всех матриц и векторов, характеризующих каждого из дикторов, составим один супервектор. Для всей совокупности дикторов эти супервекторы образуют матрицу с S столбцами. Используя метод главных компонент, найдём S собственных векторов этой матрицы. Тогда каждый новый диктор может быть представлен в виде линейной комбинации собственных векторов. Если ограничиться несколькими собственными векторами, отвечающими наибольшим собственным числам, размерность пространства можно уменьшить.

Логично было бы включать информацию о дикторах в виде вектора в пространстве дикторов в процесс обучения. Тогда адаптация заключалась бы лишь в нахождении проекции нового диктора на это пространство. Такой вид обучения называется дикторозависимым (SAT – Speaker-Aware Training). Напомним, что аббревиатуру SAT мы уже встречали в разделе 9.6, стр. 90. Там она обозначала обучение, адаптивное к диктору – Speaker Adaptive Training. Отличие этих методов состоит в том, что при адаптивном обучении информация о дикторе теряется, и каждый новый диктор приводится к некоторому среднему. При дикторозависимом обучении информация о дикторе включается во входной вектор нейронной сети (рис. 12.2.).

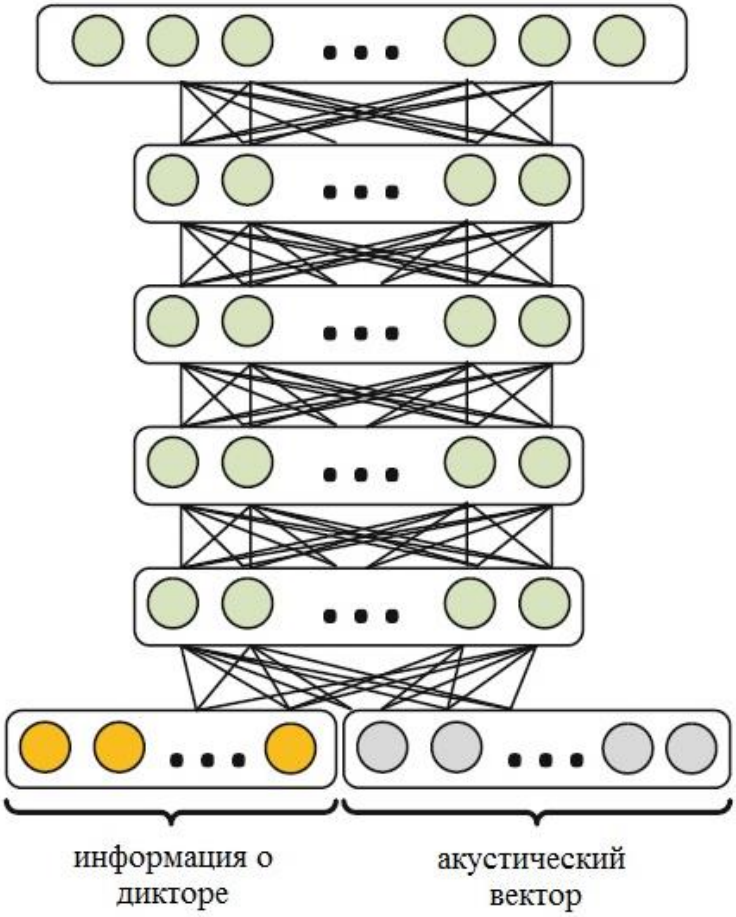


Рис. 12.2 Дикторозависимое обучение и распознавание [55].

Отметим, что наряду с дикторозависимым обучением и распознаванием, можно использовать такой же подход к адаптации к шуму и каналу связи. Такие системы называются NAT (Noise-Aware Training) и DAT (Device-Aware Training).

Преимущество дикторозависимого обучения и распознавания заключается в том, что процесс адаптации явно включён в алгоритм и не требует отдельной процедуры. Единственная проблема – надёжная оценка информации о дикторе (рис. 12.2.). Современным методом введения информации о дикторе является использование i-векторов (identity vector). Формализм i-векторов изучается в курсе по идентификации и верификации дикторов. Важной особенностью этого метода с точки зрения распознавания дикторов является то, что оценка i-вектора, характеризующего диктора, осуществляется независимо от обучения нейронных сетей, лишь на основе статистических особенностей речевой выборки для данного диктора. Ещё одной полезной особенностью является то, что размерность вектора может выбираться произвольно, в зависимости от размеров речевой выборки. Так, в работе [99], где адаптационная выборка достаточно велика, размерность i-вектора составляет 400, в работе [100] с небольшой выборкой размерность i-вектора составляет 50.

13. МОДЕЛИ ЯЗЫКА

Вернёмся к формуле (10.2). На её основании можно сделать вывод, что задачу распознавания речи можно разбить на три. Первая из них имеет дело с анализом речевого сигнала, выделением и моделированием акустических признаков. Вторая отражает зависимости, существующие между словами в языке и определяющими возможные схемы следования слов друг за другом. Наконец, задачи третьей группы связаны с определением наилучшего кандидата на распознавание среди всех возможных с использованием той информации, которая создается в ходе решения задач первых двух групп. На основании такого разделения образуются три основных модуля любой системы распознавания слитной речи: *акустическая модель*, *модель языка* и *декодер*. До сих пор рассматривались вопросы, относящиеся к первой задаче.

Основным понятием лингвистического описания речи является понятие модели языка. Произвольная модель языка позволяет формально описать язык, а точнее, те из его аспектов, которые необходимы для повышения качества автоматического распознавания речи. Определяя возможную последовательность слов, мы поднимаемся на более высокие уровни описания языка по сравнению с фонетическим и, как следствие, должны учитывать системные отношения высших порядков. Используемая модель описания слова в предложении может быть сложной, учитывающей синтаксическую и семантическую структуру высказывания, а может быть очень простой, полагающей, что появление любых слов равновероятно (в таком случае мы, по сути, отказываемся от лингвистического анализа и учета закономерностей и особенностей естественного языка).

Языковая модель – обязательная часть систем распознавания слитной речи. Не любая последовательность слов является предложением (в особенности для языков типа немецкого – с жёстким порядком слов), между словами есть грамматические и семантические связи.

Языковая модель позволяет узнать, какие последовательности слов в языке более вероятны, а какие менее. В однопроходных декодерах обычно информация от языковой модели учитывается одновременно с информацией от акустической модели (каждая со своим весом), в двухпроходных декодерах языковая модель обычно включается на втором этапе.

Использование языковой модели помогает сократить пространство поиска и снять неоднозначность при выборе из нескольких близких по стоимости акустических гипотез (для русского языка, например, помогает правильно распознать слово в нужном падеже).

Общепринятой мерой оценки моделей языка в отрыве от акустической модели является перплексия (или коэффициент неопределенности – perplexity), которая соответствует среднему коэффициенту ветвления после каждого слова, согласно модели языка. Перплексия представляет собой меру способности модели предсказывать неизвестные последовательности слов, является функцией кросс-энтропии и вычисляется по формуле:

$$PP = 2^{H(\mathbf{W})} = 2^{-\frac{1}{m}\log_2 P(w_1 w_2 \ldots w_m)} = P(w_1 w_2 \ldots w_m)^{-\frac{1}{m}}, \quad (13.1)$$

где H – кросс-энтропия текста, m – количество слов в тексте, а $P(w_1 w_2 \ldots w_m)$ – вероятность, приписываемая тексту моделью языка.

Для N-граммы вероятность $P(w_1 w_2 \ldots w_m)$ рассчитывается как произведение вероятностей всех встреченных в базе данных последовательностей N слов во фразах:

$$P(w_1 w_2 \ldots w_m) = \prod_{i=1}^{m} P(w_i \mid w_{i-1}, \ldots, w_{i-N+1}), \quad (13.2)$$

при этом для слов, начинающих фразу, вводится специальный «токен» (метка) <s>, обозначающий начало фразы. Для первых слов во фразе вероятности последовательностей рассчитываются так:

$$P(w_1 \mid <s>), P(w_2 \mid w_1, <s>), \ldots, P(w_{N-1} \mid w_{N-2}, \ldots, w_1, <s>) \quad (13.3)$$

Аналогично учитываются токены в конце фразы:

$$P(<'s>' \mid w_K, \ldots, w_{K-N-1}), P(<'s>' \mid w_K, \ldots, w_{K-N-2}), \ldots, P(<'s>' \mid w_K) \quad (13.4)$$

где K – количество слов во фразе.

Чем ниже перплексия, тем лучше модель языка. Несмотря на то, что прямой зависимости между уменьшением перплексии и улучшением качества распознавания нет, уменьшение перплексии больше, чем на 10% обычно отражается и на качестве распознавания. Очевидно, что для модели языка, где все слова равновероятны и вероятность появления слов не зависит от окружения, перплексия равна размеру словаря. По мере учёта зависимостей между словами, перплексия уменьшается до некоторого предела.

Еще одной характеристикой модели языка является процент незнакомых (внесловарных) слов (OOV – Out of Vocabulary). Он говорит о том, сколько слов из тестового текста не были найдены в произносительном словаре системы и, следовательно, не имеют никаких шансов быть распознанными. (Методы обработки OOV слов будут рассмотрены в разделе 15).

13.1. Использование условных вероятностей

Формула вычисления априорной вероятности всех N слов в предложении по полному левому контексту (то есть, на основании уже распознанных слов) может быть разложена на произведение условных вероятностей:

$$P(W) = P(w_1 w_2 \ldots w_N) = \prod_{i=1}^{N} P(w_i \mid w_1 w_2 \ldots w_{i-1}). \quad (13.5)$$

Понятно, что вычислить такую вероятность практически невозможно – это потребовало бы огромных вычислительных мощностей и невообразимого размера тренировочного корпуса. Вычисление этой вероятности потребовало бы оценки и хранения $V^{i-1}(V-1)$ независимых параметров (где V – объем словаря), а поскольку словарь хорошей системы распознавания содержит сотни тысяч слов, i может достигать значения в 100 слов и более, это становится попросту невозможным. Однако, мы можем упростить задачу, исследуя только n ближайших слов левого контекста, придя, таким образом, к понятию N-грамм:

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) = P(w_i \mid w_{i-1} \ldots w_{i-n+1}) = \frac{C(w_i \ldots w_{i-n+1})}{C(w_{i-1} \ldots w_{i-n+1})}, \quad (13.6)$$

где C – количество соответствующих N-грамм (последовательностей соседних n-слов).

Подсчитав вероятности N-грамм на основании большого тренировочного корпуса (который может достигать сотен миллионов словоформ), мы можем в дальнейшем использовать их для определения вероятности следования слов друг за другом в незнакомом тексте (представляющим собой последовательность слов – кандидатов на распознавание). Такой подход называется оценкой максимального правдоподобия (MLE – Maximum Likelihood Estimation). В моделях языка для систем распознавания обычно используются N-граммы порядка 2 и 3, называемые биграммами и триграммами.

Данные модели (не только биграммные и триграммные, а вообще все модели, учитывающие ограниченный контекст) имеют два серьезных и сразу бросающихся в глаза недостатка. Во-первых, подобное описание часто оказывается лингвистически недостоверным в связи с тем, что не учитывает релевантные слова, если они находятся на расстоянии больше N от исследуемого слова, но в то же время учитывает слова с низкой предсказательной способностью, которые просто оказались ближе (например, вследствие своей высокой частотности). Во-вторых, при желании учесть больший контекст, оказывается, что для этого потребуется очень серьезное увеличение тренировочного корпуса, и он может принять поистине огромный размер – что говорит о том, что тренировочный корпус используется неэффективно.

13.2. Статистическое сглаживание

Методы статистического сглаживания позволяют учитывать факты появления N-грамм, которые ни разу не встречались в тренировочном корпусе. Действительно, нелогично полностью отказываться от гипотезы, которая обладает очень большой акустической вероятностью только на основании того, что одна из N-грамм в гипотезе не встретилась ни разу в тренировочном корпусе. Этот факт, скорее, говорит о том, что тренировочный корпус не может охватить все возможные N-граммы языка, а не о том, что данная N-грамма невозможна в языке. Для решения этой проблемы используются алгоритмы *сглаживания* (smoothing) и *отката* (back-off). Идея сглаживания заключается в том, что мы берем часть вероятностной массы у встретившихся в корпусе N-грамм, и распределяем ее по всем возможным комбинациям слов из словаря, составляющим множество невстреченных (unseen) N-грамм. Известными алгоритмами сглаживания являются пересчет Гуда-Тьюринга, Виттена-Белла, сглаживание Кнезера-Нея. Откат подразумевает использование вероятностей N-1-грамм в том случае, если соответствующая N-грамма имеет нулевую вероятность.

Сглаживание и откат являются обязательными элементами практически любой модели языка. Существует большое число различных алгоритмов сглаживания, однако новые разновидности лишь очень незначительно улучшают

результаты, достигаемые с алгоритмами, упомянутыми выше. В принципе, можно сказать, что алгоритмы сглаживания достигли предела своей эффективности, и дальнейшие усилия по их усовершенствованию не представляются перспективными.

Кроме сглаживания и отката существует ряд других дополнений стандартной N-граммной модели, которые являются опциональными и могут использоваться в зависимости от конкретного языка.

13.3. Классовые модели

Одним из наиболее эффективных усовершенствований N-граммной модели является использование информации о принадлежности слов к тем или иным классам эквивалентности. Особенно важным это оказывается, если лингвистическая модель строится на относительно небольшом тренировочном корпусе. В таком случае может оказаться, что при использовании системы будет встречаться так много N-грамм с нулевым значением (не встретившихся в тренировочном корпусе), что даже сложные алгоритмы сглаживания не смогут обеспечить нужного приближения. Использование N-грамм, членами которых вместо конкретных слов являются классы, к которым принадлежат слова, позволяет решить эту проблему. Это связано с тем, что количество классов, в которые можно объединять слова, несравнимо меньше количества слов, следовательно, можно быть уверенными, что даже при небольшом тренировочном корпусе охват всех возможных (классовых) N-грамм будет достаточно полным.

Обычно условная вероятность слова в чистом виде определяется как произведение вероятности класса на основе предыдущих классов и вероятности того, что наше слово принадлежит данному классу

$$P(w_n \mid w_{n-N+1}^{n-1}) = P(w_n \mid c_n) P(c_n \mid c_{n-N+1}^{n-1}), \qquad (13.7)$$

где *c* соответствует классу для каждого конкретного слова.

Данные множители могут быть вычислены по формулам

$$P(w \mid c) = \frac{C(w)}{C(c)} \quad \text{и} \quad P(c_i \mid c_{i-1}) = \frac{C(c_{i-1} c_i)}{\sum_c C(c_{i-1} c)}, \qquad (13.8)$$

где (*C(x)* – каунт единицы, то есть то, сколько раз она встретилась в тренировочном корпусе)

Как мы видим, N-граммная модель, построенная на классах слов, дополняется вероятностями принадлежности слова к классу. На основании способа отнесения слов к тем или иным классам классовые модели можно разделить на три основные категории:

- модели, работающие с классами, построенными вручную;
- модели, опирающиеся на частеречные классы слов (с использованием информации о принадлежности слов к тем или иным грамматическим категориям);
- модели классов, построенных при помощи алгоритмов статистической кластеризации.

13.4. Морфемные модели

Для синтетических языков (флективных и особенно агглютинативных языков) оказывается весьма перспективным использовать при построении модели языка информацию о морфемном составе слов (в виде традиционного морфемного членения, либо в виде членения на псевдоморфы). Для этого сначала используются процедуры автоматического определения морфемных границ, информация о которых в дальнейшем используется при построении N-граммной модели языка.

Несмотря на то, что существуют модели, предполагающие осуществление полного морфологического анализа, для славянских языков были показаны преимущества модели основа/флексия. В таком случае триграммная модель, например, определяется формулой

$$P(s_i \mid w_1 \ldots w_i) = P(s_i \mid s_{i-2} s_{i-1})$$
$$P(e_i \mid w_1 \ldots w_i) = P(e_i \mid s_i e_{i-1}), \qquad (13.9)$$

где s_i – основа i-того слова; e_i – флексия (окончание) i-того слова.

Такие модели создавались и апробировались и для русского языка [71] и [42], однако, не принесли желаемого успеха в плане повышения точности распознавания речи. Это, видимо, связано с тем, что морфемы в русском языке короткие и каждая из них имеет по несколько вариантов произнесения в зависимости от лингвистического контекста, кроме того, в русскоязычных словоформах наблюдается чередование звуков и изменение морфем (например, корневых) в зависимости от грамматических характеристик словоформ, что не позволяет легко соединять цепочки распознанных морфем в правильные слова.

Однако морфемные модели языка довольно успешно применяются для ряда агглютинативных языков, например, финского [72], турецкого [73], венгерского [74] и т.д.

13.5. Синтаксические и семантические модели

Для многих синтетических естественных языков, в том числе и русского, характерен практически свободный порядок слов в предложении. Как следствие, использование моделей языка на основе N-грамм для распознавания речи дает гораздо более низкие результаты, что объясняется необходимостью нахождения связей между словами по принципу, отличному от простого соположения.

Для таких случаев эффективным может оказаться учет синтаксических связей слов в предложениях при моделировании языка. Если мы хотим использовать синтаксическую информацию для распознавания речи, то для начала нам нужно автоматически выделить ее из предложения: таким образом, наша задача разбивается на две практически независимые задачи. Одной из первых таких моделей стала структурированная модель языка (Structured Language Model) [43, 75]. Структурированная языковая модель применяется на стадии декодирования речи для синтаксического разбора результатов распознавания, синтаксическое дерево строится динамически в ходе распознавания. Модель позволяет использовать дальнодействующие связи между

словами и предсказывать слово не только по нескольким предыдущим лексическим единицам, но также по доступным главным словам. Такая модель хорошо подходит для аналитических языков с жесткой грамматической структурой (например, для английского языка).

В [76] был предложен подход синтаксических N-грамм (SN-грамм). В случае SN-грамм соседние слова выбираются в соответствии с их синтаксическими связями в синтаксических деревьях, а не в соответствии с тем, как они появились в тексте. При этом используются традиционные N-граммы слов, а также метки и характеристики частей речи. В другой работе [77] была предложена стохастическая морфо-синтаксическая модель для системы распознавания венгерской речи, эта модель описывает допустимые в венгерском языке словоформы (комбинации морфем).

Кроме того, в [78] предложены составные языковые модели с введением понятия категорной языковой модели и, в частности, категорных n-грамм. Каждому слову в словаре приписываются 15 атрибутов, определяющих грамматические свойства словоформы. Множество значений атрибутов определяет класс словоформы. Каждое слово в предложении рассматривается как его начальная форма и морфологический класс. В итоге языковая модель разбивается на две составляющие: изменяемую часть (основанную на морфологии) и постоянную часть (основанную на начальных формах слов), которая строится как n-граммная языковая модель.

Кроме того, для того чтобы учесть дальнодействующие связи между словами во фразе, недавно была также предложена синтаксическо-статистическая модель языка системы распознавания русской речи со сверхбольшим словарем. Такая модель позволяет объединить статистический анализ (N-граммы) и синтаксический анализ обучающего текстового корпуса [79]. В данном подходе статистическая N-граммная модель языка расширена (посредством интерполяции) за счет синтаксического анализа обучающего текстового корпуса. В ходе статистического анализа русскоязычных текстов выявляются новые (потенциальные) N-граммы, содержащие грамматически связанные пары слов, которые были разделены в обучающем тексте другими словами. Синтаксический анализ позволяет увеличить количество создаваемых в результате обработки текста различных N-грамм и, тем самым, повысить качество модели языка за счет выявления грамматически связанных пар слов и снизить процент ошибок распознавания слов в слитной речи.

Область семантики (смысла слов) же при лингвистическом анализе естественного языка является пока самой нечеткой и трудноисследуемой. Из всех уровней языка именно семантическая информация сокрыта в тексте глубже всего и, как следствие, хуже всего поддается формальному описанию. Как следствие, использование семантической информации в системах распознавания речи на данный момент весьма ограничено. Задача выделения из текста достоверной семантической информации (глубокий семантический анализ) является пока нерешенной.

13.6. Модели темы высказывания

Известно, что хорошее статистическое описание языка в целом требует наличия не только очень большого тренировочного корпуса, но и максимально широкого охвата текстов различных жанров и стилей. С другой стороны, если для определенного тестового текста мы будем знать его принадлежность к какому-либо классу (например, экономика, юриспруденция и т.п.), то использование для такого текста модели, натренированной на текстах соответствующего класса, приведет к серьезному улучшению качества распознавания. По сути, мы используем различные модели языка для текстов с различной темой. Основа модели остается одной и той же, но распределение вероятностей оказывается различным. Поэтому дополнительной задачей является автоматическое выделение набора тем для тренировочного корпуса и тренировка модели для того, чтобы затем иметь возможность относить исследуемый текст к одной из таких тем.

13.7 Модели языка на основе нейронных сетей

Во всех рассмотренных выше моделях языка слова представлены в дискретном пространстве, а именно в словаре. Такое представление не позволяет выявить «похожесть» слов и оценить вероятность не встреченных в базе N-грамм, поскольку рассмотренные выше процедуры отката и сглаживания просто не позволяют присвоить не встреченной N-грамме нулевую вероятность. Наглядный пример, приведённый в работе [101], иллюстрирует, что понимается под похожестью. Рассмотрим две фразы: «кошка идёт по комнате» и «собака вбежала в кухню». Очевидно, что пары слов «кошка - собака», «комнате – кухню» и предлогов «по – к» имеют сходные семантические и синтаксические значения. Идея заключается в том, чтобы организовать непрерывное пространство, в котором подобные похожие слова были представлены в каком-то смысле близкими друг к другу векторами. Тогда, даже если некоторое сочетание слов не было встречено в обучающей выборке, система сможет восполнить недостаток знаний по аналогии с присутствующими в базе примерами и оценить его вероятность.

На рисунке 13.1 представлена нейронная сеть с двумя скрытыми слоями, реализующая модель языка в непрерывном пространстве [101,102,103].

На вход подаётся супервектор, составленный из векторов слов контекста (на рисунке контекст состоит из трёх слов). Вектор, соответствующий слову строится по принципу 1-из-N, то есть все элементы вектора, кроме i-го равны нулю, элемент с номером i равен 1, где i – номер слова в словаре, N – размер словаря.

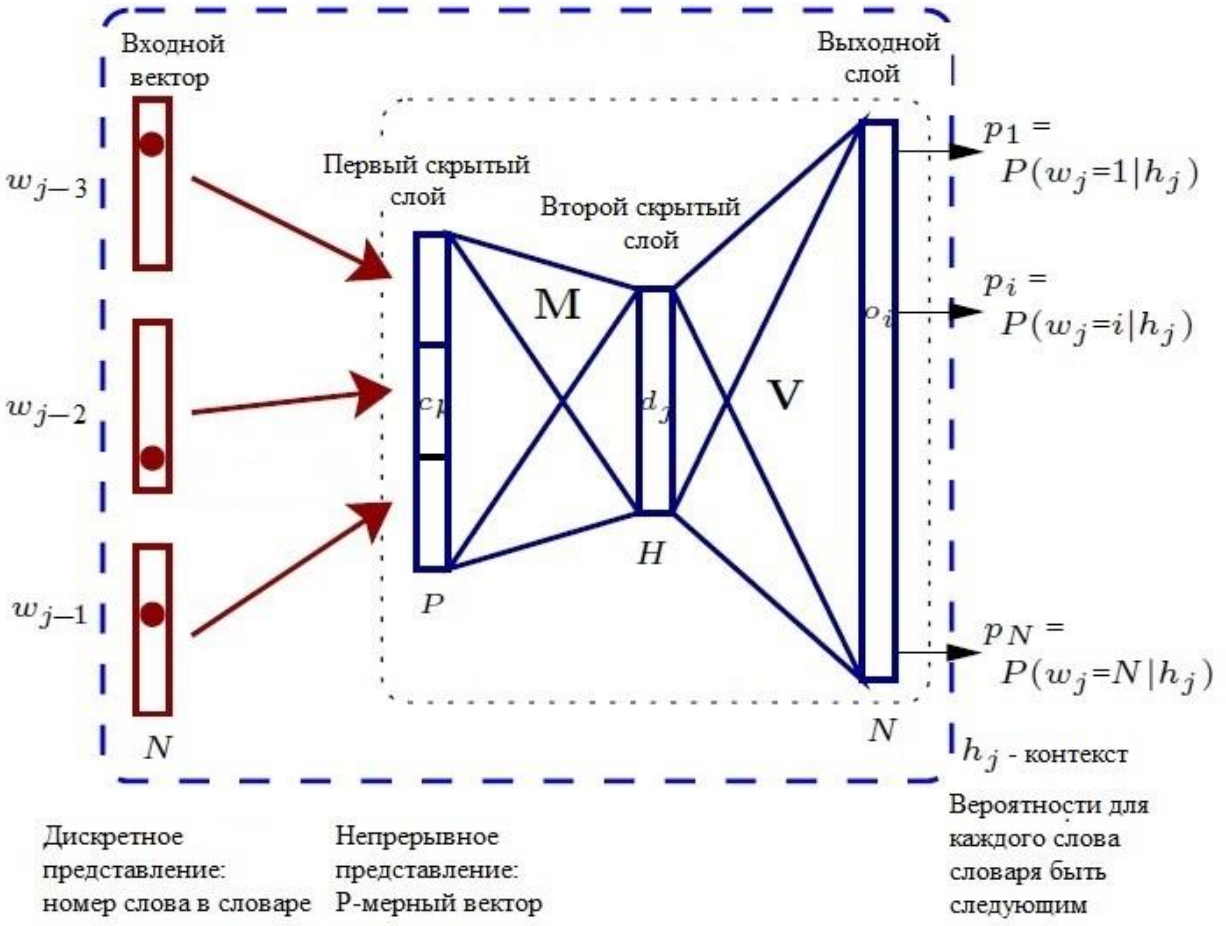


Рис. 13.1 Модель языка на основе нейронной сети [102].

Функция активации первого скрытого слоя линейна, то есть элементы состояния первого скрытого слоя c_j представляют собой линейные комбинации со смещением компонент входного супервектора.

Состояния второго скрытого d_j и выходного слоя o_j вычисляются следующим образом:

$$d_j = \tanh(\sum_l m_{jl} c_l + b_j) \quad , \tag{13.10}$$

$$o_i = \sum_j v_{ij} d_j + k_i \, , \tag{13.11}$$

где матрицы m_{jl} и v_{ij} – весовые матрицы соответствующих слоёв, b_j и k_i – элементы векторов смещения.

Вероятность для каждого слова словаря быть следующим в этом контексте вычисляется обычным образом по формуле softmax:

$$p_i = \exp(o_i) / \sum_{k=1}^{N} \exp(o_k) \, , \tag{13.12}$$

где N – размер словаря и выходного слоя.

Отметим, что без первого линейного скрытого слоя можно обойтись, включив его матрицу преобразования во второй скрытый слой. Для этого надо вместо элементов c_l подставить их значения, полученные в результате линейного преобразования из входного вектора и поменять порядок суммирования в

формуле (13.10). Это замечание сделано для того, чтобы отсутствие линейного слоя в рекуррентной модели (см. ниже) не казалось недостатком.

Размер первого скрытого слоя обычно выбирается в диапазоне 50 – 200, второго – 400 – 1000, , в зависимости от размеров обучающей выборки. Модель исследовалась на словарях размером от 40000 до 200000 слов.

Использование этой модели языка позволило уменьшить перплексию с 70.2 до 67.6 и уменьшить ошибки распознавания слов с 14.24% до 14.02% и далее до 13.92% при увеличении размеров второго скрытого слоя с 400 до 1000 [102].

Явным недостатком рассмотренной модели языка является фиксированная длина контекста. Очевидно, что в языке существуют довольно устойчивые сочетания слов различной длины. Рекуррентная нейронная сеть лишена этого недостатка, поскольку может хранить информацию о контекстах произвольной длины [104] (Рис. 13.2).

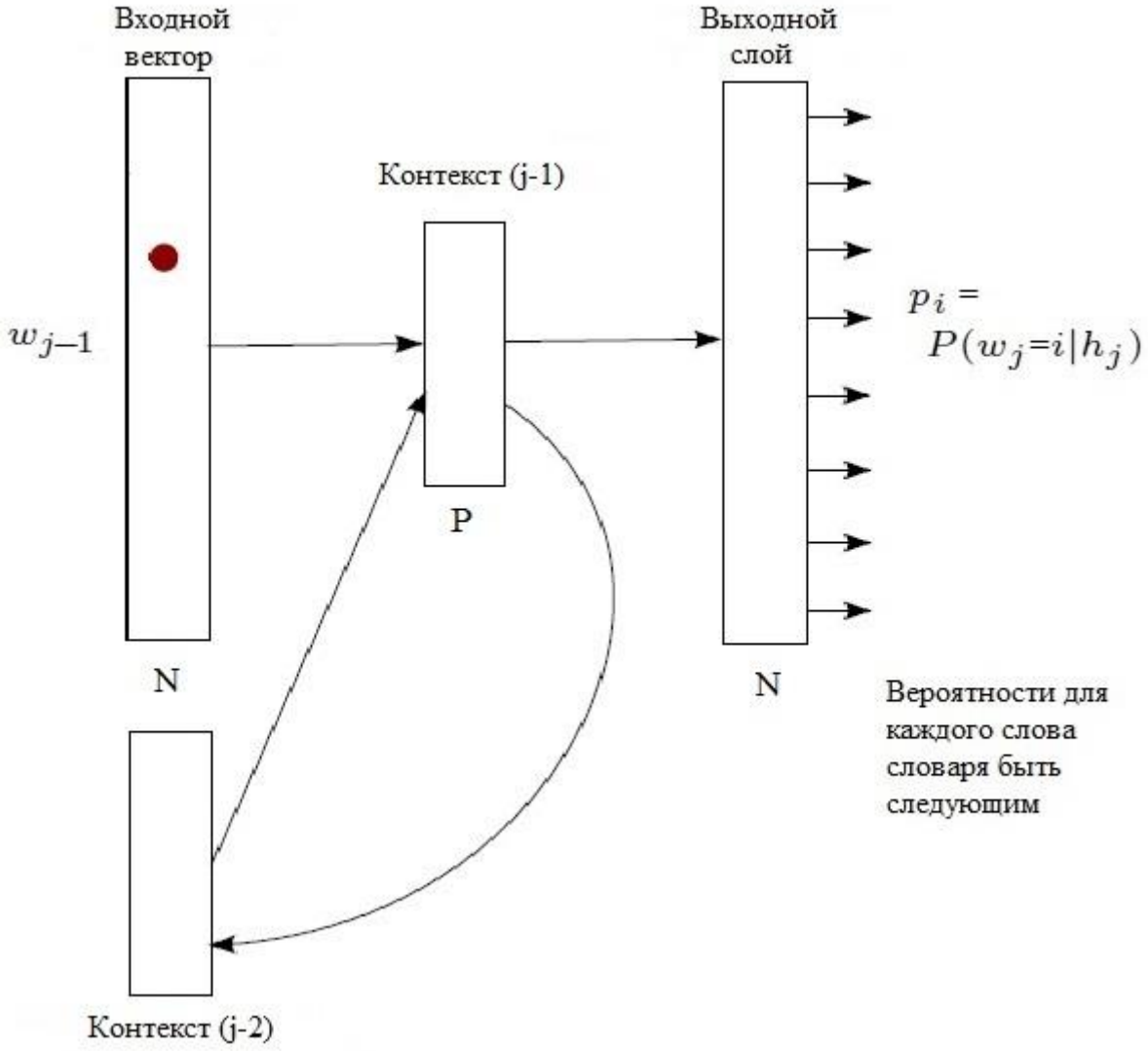


Рис. 13.2 Модель языка на основе рекуррентной нейронной сети [104].

В качестве входного вектора используется конкатенация 1-из-N очередного вектора слова и вектора состояния скрытого слоя нейронной сети на предыдущем шаге. Состояние скрытого слоя отражает информацию о накопленном контексте.

Состояния скрытого d_j и выходного слоя o_j, а также вероятности слов вычисляются так же, как для нерекуррентной сети (13.10, 13.11, 13.12), но вместо активационной функции *tanh* используется сигмоидальная функция, что не принципиально. Также отсутствуют смещения.

Модель языка исследовалась со следующими параметрами: размерность скрытого, контекстного слоя от 30 до 500, размер словаря N – от 30000 до 200000 слов. Отметим, что в отличие от языковых моделей, основанных на нерекуррентных нейронных сетях [101,102,103], модель языка, основанная на рекуррентной нейронной сети, имеет только один параметр, который надо выбрать заранее – размер контекстного слоя.

Рассмотренная модель языка показала явное преимущество перед моделями сглаживания Кнезера-Нея (раздел 13.2) – количество ошибок распознавания слов уменьшилось на 18%, при этом база обучения составляла всего 6.4 миллиона слов вместо 37 миллионов в сравниваемой системе. Надо отметить, что меньшая по объёму база обучения использовалась не случайно – процедуры обучения рекуррентных нейронных сетей чрезвычайно требовательны к производительности компьютеров. Обучить такую систему на имеющихся базах, содержащих более миллиарда слов, в настоящее время можно только на суперкомпьютерах. Таким образом, ещё раз, как и в разделе 12.2, отметим необходимость разработки новых процедур обучения рекуррентных нейронных сетей.

14. ДЕКОДЕР

В ходе работы системы автоматического распознавания речи задача распознавания сводится к определению наиболее вероятной последовательности слов, соответствующих содержанию речевого сигнала. Наиболее вероятный кандидат должен определяться с учетом как акустической, так и лингвистической информации. Это означает, что необходимо производить эффективный поиск среди возможных кандидатов с учетом различной вероятностной информации. При распознавании слитной речи число таких кандидатов огромно, и даже использование самых простых моделей приводит к серьезным проблемам, связанным с быстродействием и памятью систем. Как результат, эта задача выносится в отдельный модуль системы автоматического распознавания речи, называемый декодером.

Декодер должен определять наиболее грамматически вероятную гипотезу для неизвестного высказывания – то есть определять наиболее вероятный путь по сети распознавания, состоящей из моделей слов (которые, в свою очередь, формируются из моделей отдельных фонов). Правдоподобие (likelihood) гипотезы определяется двумя факторами, а именно вероятностями последовательности фонов, приписываемыми акустической моделью, и вероятностями следования слов друг за другом, определяемыми моделью языка.

В случае распознавания слитной речи сеть вариантов распознавания оказывается настолько большой, что исследование и оценка всех возможных вариантов представляется невыполнимой с вычислительной точки зрения в режиме, сопоставимым с режимом реального времени. В связи с этим, оказывается необходимой разработка как можно более эффективных алгоритмов быстрого поиска, которые уже не будут гарантировать нахождение оптимального варианта распознавания, однако будут осуществлять поиск за приемлемое время. Как правило, поиск осуществляется в пределах «луча» (Beam Search), то есть все гипотезы, вероятность или правдоподобие которых уступает лучшей гипотезе на величину, превышающую некоторый порог, отбрасываются. Очевидно, что чем чаще производится анализ гипотез и чем раньше лишние гипотезы отбрасываются, тем эффективнее будет работа декодера, поскольку количество гипотез на каждом узле веерообразно увеличивается, напоминая лавину. Ещё одним способом ограничения требований к памяти и быстродействию является сохранение N лучших гипотез.

Рассмотрим математическую основу декодеров. Отбрасывая несущественный на этапе распознавания знаменатель, перепишем (10.2):

$$W = \arg\max_{W} P(W)P(X|W), \qquad (14.1)$$

где $X = x_1^T = x_1,\ldots,x_N$ – последовательность векторов признаков входного сигнала, $W = w_1^n = w_1,\ldots,w_n$ – последовательность слов, принадлежащих словарю размером N_W.

Первый множитель *P(W)* описывает вклад лингвистического модуля, второй *P(X|W)* – лексического, фонетического и акустического источников знаний. В

соответствии с концепцией марковских цепей, второй множитель представляет собой сумму вероятностей всех возможных последовательностей состояний, что приводит к уравнению:

$$\hat{W} = \arg\max_{W}\{P(W)\cdot\sum_{s_1^T} P(x_1^T, s_1^T \mid w_1^N)\}, \qquad (14.2)$$

где s_1^T – одна из последовательностей состояний, порождаемых последовательностью слов w_1^n. На практике применяется критерий Витерби – ищется последовательность состояний, дающая максимальный вклад в сумму (14.2):

$$\hat{W} = \arg\max_{W}\{P(W)^{\alpha}\cdot \underset{s_1^T}{Max}[P(x_1^T, s_1^T \mid w_1^N)]\}\ . \qquad (14.3)$$

Отметим, что в формулу (14.3) введён эмпирический параметр α, оптимизируемый по обучающей выборке.

Сложность декодирования заключается в комбинаторном характере задачи и, следовательно, огромном числе переборов при полном решении, что требует внедрения эффективных эвристик для практического решения задачи. Структура сети, на которой осуществляется поиск, является продуктом ограничений, налагаемых на неё источниками знаний на каждом уровне (состояний, фонем, слов) как внутри слов, так и между ними. На основании предыдущего материала очевидно, что диапазон влияния этих ограничений характеризуется «короткой памятью». Этот факт позволяет оптимизировать процесс поиска на основании принципа ранней рекомбинации, который формулируется следующим образом:

если несколько гипотез в сети имеют общий узел, следует оставить наилучшую гипотезу до этого узла и отбросить остальные, поскольку при дальнейшем развитии процесса у этих гипотез уже не будет возможности превзойти сохранённую

Использование N-граммной модели языка имеет два явных последствия:

А. сеть не имеет ограничений на ветвления – за любым словом может следовать любое другое;

Б. вероятности слов зависят только от *N*-1 предшествующих:

$$P(w_n \mid w_{n-1}, w_{n-2}, ..., w_1) = P(w_n \mid w_{n-1}, w_{n-2}, ..., w_{n-N+1})\,. \qquad (12.4)$$

Отсюда следует, что хранить историю всех гипотез, простирающуюся далее *N*-1 слов, не имеет смысла, однако удаление гипотез на более ранних этапах может привести к ошибкам. Способ сохранения, анализа и обработки гипотез является предметом исследований при создании декодеров.

14.1. Организация лексикона в виде префиксного дерева

Лексикон (словарь) определяет список слов (обычно это линейный список) с их фонетической транскрипцией в виде небольшого количества контекстно-независимых (не учитывающих коартикуляцию) фонетических символов. Некоторые слова могут иметь варианты произнесения с приписанными им вероятностями.

Представление лексикона в виде префиксного дерева обеспечивает более компактное описание с уменьшенным количеством дуг, особенно на начальных участках слов, куда приходятся основные вычислительные затраты при поиске. Поскольку любое слово лексикона может следовать за любым другим, объединение одинаковых фрагментов деревьев сокращает количество дуг, которые необходимо рассмотреть для того, чтобы генерировать стартовые гипотезы следующего слова.

Одной из проблем, связанных с использованием префиксных деревьев, является то, что вероятность слова, получаемая из модели языка, определяется только в последнем узле и не оказывает влияние на формирование гипотез до этого момента. В качестве выхода предлагается распределять вероятность слова по дугам его префиксного дерева.

Префиксное дерево может быть построено на основе контекстно-независимых фонемных транскрипций, или с использованием контекстно-зависимых трифонов, что приводит к увеличению количества дуг с нескольких десятков до нескольких сотен.

Использование контекстно-зависимых межсловных трифонов ещё более усложняет задачу – в этом случае последняя дуга предыдущего слова расщепляется на множество дуг, поскольку она теперь зависит от первого трифона следующего слова.

Расширение сети, осуществляемое до декодирования (Static network expansion), было естественным решением на начальном этапе развития систем распознавания речи. С увеличением объёма словаря и усложнением используемых источников знаний использование этого метода становится всё более затруднительным. Временным выходом является либо переход к динамическому расширению сети (Dynamic search network expansion), либо поиск резервов уменьшения размеров сети. В частности, используется то, что реальное количество узлов и дуг существенно (на несколько порядков) меньше, чем теоретически возможное, поскольку не все сочетания монофонов в виде бифонов или трифонов встречаются в языке, а также то, что начальные и конечные состояния различных трифонов могут связываться (заменяться одним состоянием).

14.2. Использование взвешенных конечных автоматов

Для декодирования используется также аппарат взвешенных конечных автоматов (Weighted finite automata). На рис. 12.1. представлен простой пример такого автомата.

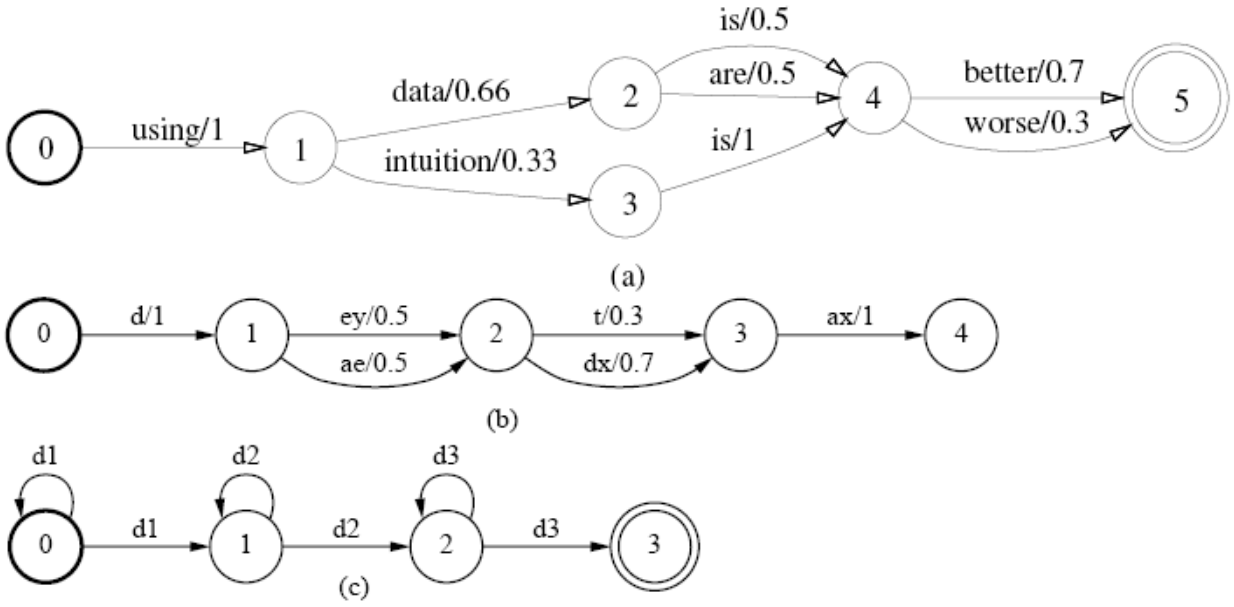


Рис. 12.1. Взвешенный конечный автомат. Метки l и веса w перехода обозначены на соответствующей направленной дуге как l/w [44].

Разрешённые слова с их вероятностями проставлены на дугах, составляющих пути. Каждое слово представляется также автоматом. На рис. 12.1(б) изображён граф слова “data”, с двумя вариантами произнесения. На рис.12.1(в) автомат описывает стандартную марковскую цепь для фонемы ‘d’. Полная вероятность произнесения подсчитывается как произведение вероятностей, полученных на всех вложенных автоматах. Эти автоматы состоят из набора промежуточных состояний, начального состояния и набора конечных состояний, соединённых переходами. Каждый переход имеет начальное и конечное состояния, метку и вес. Задачей декодера является оптимизация автомата. Декодер находит в лексиконе варианты произнесений слов и подставляет их в грамматику. Представление в виде фонетического дерева на данном этапе может быть использовано для уменьшения количества путей. Далее декодер определяет контекстно-зависимые модели для каждой фонемы в контексте и подставляет их в граф.

14.3. Использование взвешенных преобразователей с конечным числом состояний

В последние годы получил развитие подход к статическому декодированию, основанный на «взвешенных преобразователях с конечным числом состояний» (WFST – Weighted Finite State Transducer) [44]. Образец WFST с тем же лексиконом, что и на рис. 12.1., изображён на рис. 12.2.

Каждый переход на рис. 12.2(б) имеет идентичные метки входа и выхода. Поскольку слова кодируются выходной меткой, стало возможно объединять преобразователи для нескольких слов (слова data и dew на рис. 12.2(б). Аналогично можно объединять марковские модели фонем. Это иллюстрирует основное преимущество преобразователей над взвешенными автоматами:

преобразователи могут объединять различные уровни представления, например, уровень фонем и уровень слов.

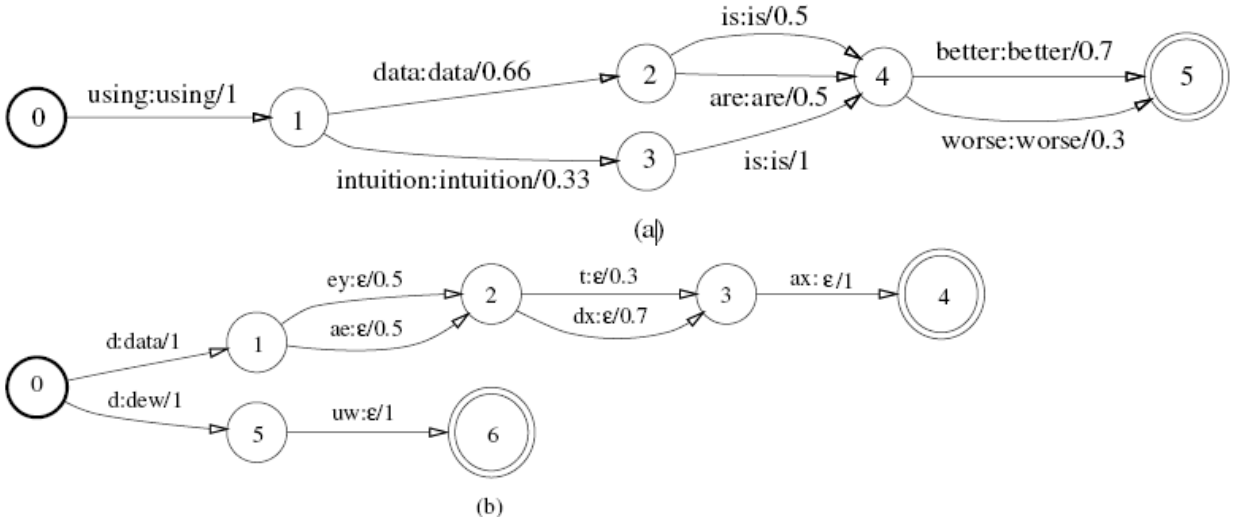


Рис. 12.2. Пример WFST [44].

Благодаря этому подходу, удаётся объединить в одну сеть WFST различные источники знаний – марковские модели, лексиконы, N-граммные статистические модели языка. В данной сети входными метками являются состояния марковских моделей, выходными – слова. Разработанные в теории конечных автоматов методы детерминизации и минимизации позволяют получить чрезвычайно компактные сети. Отметим, что при поиске оптимального пути на графе декодеру не придётся обращаться к представлению фонем, лексикону, модели языка – вся информация уже заключена в структуре графа. Благодаря этому, декодер упрощается и ускоряется – декодеру остаётся только подставлять вероятности эмиссии в соответствии с рассматриваемыми гипотезами.

Метод декодирования, в котором расширение сети интегрировано в процесс декодирования, называется «динамическим расширением сети» (Dynamic search network expansion). Построенное на старте распознавания начальное дерево расширяется, используя виртуальные узлы и временные структуры, содержащие только информацию, необходимую для текущей гипотезы. Разновидность метода динамически расширяемой сети, называемая синхронным по времени методом, особенно удобна для анализа вероятностей гипотез в каждый момент времени, поскольку позволяет отбрасывать гипотезы, не попавшие в луч, не дожидаясь концов слов.

15. ПРОБЛЕМА ВНЕСЛОВАРНЫХ СЛОВ

Существующие системы распознавания слитной речи содержат модели десятков и сотен тысяч слов, однако, как и при построении моделей фонем (см. раздел 8), никакие базы данных не могут обеспечить полное покрытие словаря в условиях реальной эксплуатации. Понятно, что если не предусмотреть способов обработки таких случаев, внесловарное слово, или OOV-слово (Out Of Vocabulary) будет распознано, как одно из слов словаря – IV-слово (IV – In-Vocabulary). Причём такая вставка в текст может вызвать цепочку дополнительных ошибок. Поясним сказанное.

Согласно формулам 6.2, 6.4, вероятность того, что наблюдения *X* порождены моделью *W* и вероятность того, что модели *W* соответствуют наблюдения *X*, связаны соотношением:

$$P(W \mid X) = P(W)P(X \mid W), \qquad (15.1)$$

где *P(W)* – вероятность модели, то есть произведение вероятностей слов данной последовательности, представленной последовательностью моделей фонем.

Таким образом, слова не распознаются последовательно, одно за другим – решение откладывается до момента распознавания последнего слова в цепочке, при этом сохраняется несколько вариантов цепочек (гипотез). То, что слово не включено в словарь, означает, что его априорная вероятность равна нулю, и его участие в любой гипотезе исключено – произведение вероятностей тоже будет равно нулю. Вместо OOV-слова будет подставлено некое созвучное слово, либо несколько коротких IV-слов. Поскольку сочетания слов в моделях языка имеют определённые вероятности, ошибка может распространиться на соседние слова. Стоит отметить также, что неправильное распознавание OOV-слов может приводить к моделям с низкими вероятностями, что приводит к необходимости увеличивать количество рассматриваемых гипотез, что, в свою очередь, увеличивает объём вычислений. Учитывая, что системы распознавания с большими словарями работают на пределе вычислительных возможностей существующих компьютеров, такой сценарий очень нежелателен.

Какие же слова могут оказаться внесловарными? Анализ показывает, что наибольшую долю среди OOV-слов занимают новые термины, имена, названия. Это как раз те слова, которые чаще всего определяют смысл высказывания, то есть, собственно, те слова, ради которых фраза и была произнесена. Иначе говоря, OOV-слова могут нести большой объём информации.

Из вышесказанного следует, что задача обработки OOV-слов очень важна и должна включать следующие подзадачи:

1. определение наличия и положения слова во фразе;
2. распознавание последовательности фонетических единиц, составляющих слово;
3. определение написания слова (рассматриваться не будет).

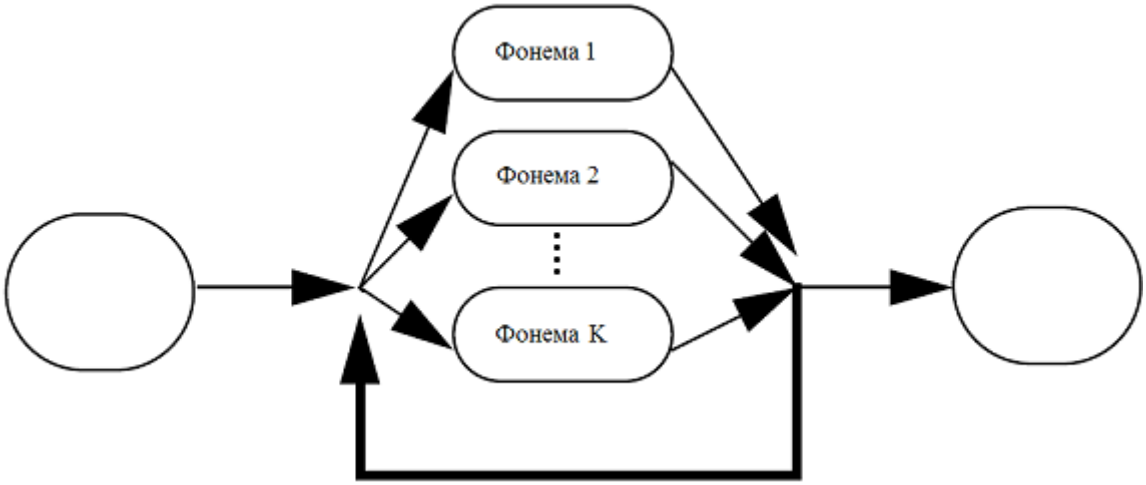


Рис. 15.1. Фонемный граф общей модели слова.

Решать проблему OOV-слов можно несколькими способами, или их комбинациями:

- Увеличение размера словаря – очевидный метод, не нуждающийся в пояснениях.
- Введение общей модели слова в словарь – расширение идеи моделей заполнения (Filler models), или моделей мусора (garbage) и моделей неречевых звуков.
- Использование системы с двумя фазами распознавания, на первой из которых распознаются более крупные, чем фонемы, единицы – Sub-Word Units (например, слоги, или полученные автоматически сочетания фонем).
- Использование доверительных оценок, полученных различными системами распознавания.

15.1. Использование моделей заполнения

Одним из методов обработки OOV-слов является использование моделей заполнения (Filler models) [2]. Модель заполнения представляет собой универсальную модель слова, в которой возможны произвольные последовательности фонем (рис. 15.1). Эта модель подставляется в словный граф (рис. 15.2), как обычное слово, но вход в неё затрудняется некоторым порогом, который подбирается эмпирически на большом речевом материале, содержащем слова, не входящие в словарь системы.

Языковая модель, учитывающая OOV-слова, строится на основе очень большой текстовой базы данных и включает OOV-слова в n-граммы. С такой языковой моделью можно дать доверительную оценку различных гипотез, включающих и не включающих OOV-слова. Вероятность OOV-слова является произведением вероятностей фонем при прохождении по графу общей модели слова:

$p = p_1 p_2 \cdots p_N$, где N – количество фонем.

Тогда, с учётом (14.1), решение о наличии OOV-слова принимается, если

$$P(X \mid p)P(p \mid OOV)P(OOV) > P(X \mid w_i)P(w_i) \qquad (15.2)$$

для любого слова w_i, принадлежащего словарю.

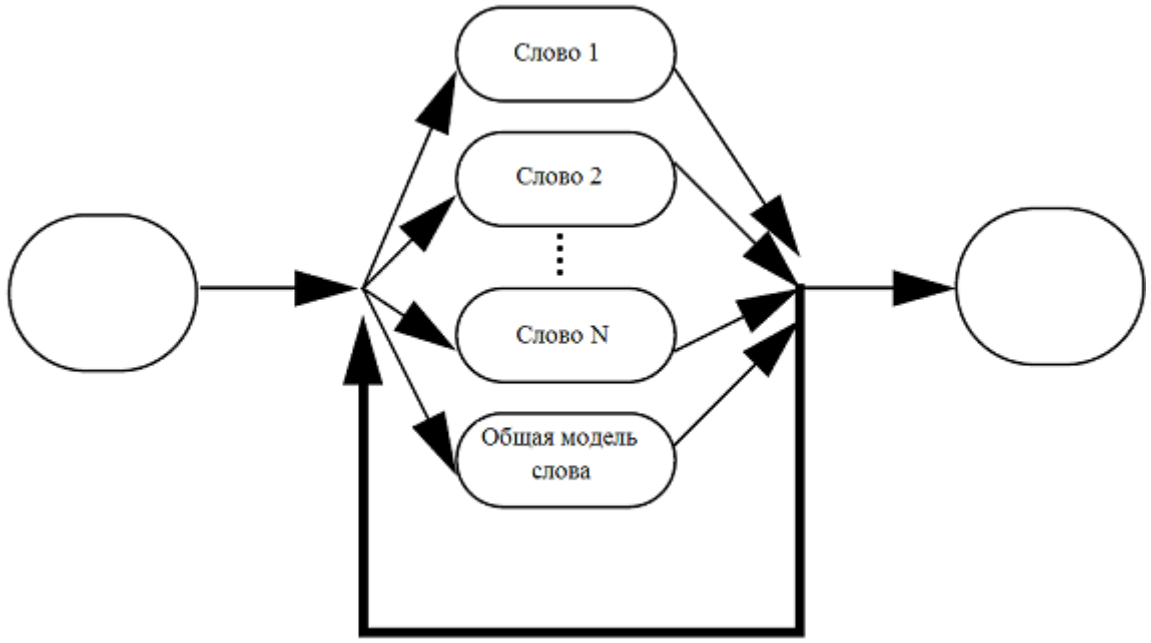


Рис. 15.2. Словный граф модели фразы.

Схема общей модели слова допускает усовершенствования: после распознавания и получения альтернативных цепочек фонем с различными вероятностями к ним применяются ограничения в виде фонемной грамматики. Грамматика определяет вероятности следования фонем и эквивалентна N-граммным моделям языка. Фонемная грамматика может определяться на том же обучающем корпусе, что и вся система, или учитывать только статистику OOV-слов (Oracle), что даёт существенное улучшение параметров системы, но, по-видимому, пригодно, в основном, для фиксированной предметной области, где фонетика OOV-слов имеет какие-то общие черты.

Отметим, что сходный приём используется для обработки так называемых «артефактов» – неречевых звуков (кашель, мычание, звуки шагов, хлопанье двери,…) и выделения «ключевых слов». В случае артефактов модель, конечно, строится не на основе моделей фонем, а представляет собой статистическую обработку соответствующего артефакта из достаточной по размеру базы данных артефактов. В задаче выделения ключевых слов для принятия решения вероятность появления произвольного слова сравнивается с вероятностью ключевого слова с наибольшей вероятностью из списка.

15.2. Использование фиксированных комбинаций фонем

Использование фиксированных комбинаций фонем, или частей слов (Sub-Word Units - SWU) [48] является развитием идеи ограничений на возможные фонемные цепочки. Понятно, что фиксированный набор некоторых комбинаций

фонем ограничивает возможные цепочки существеннее, чем грамматики, которые разрешают любые комбинации, хоть и с различными вероятностями.

Для построения SWU используется уже знакомый по разделу 8.2 метод, управляемый данными (data-driven approach). Оценивается статистика сочетаний фонем или их комбинаций на большой базе данных. Процесс носит итеративный характер и начинается с отдельных фонем. Критерием для объединения некоторых комбинаций фонем в новую комбинацию является взвешенная взаимная информация MI_w двух комбинаций:

$$MI_w(u_i,u_j) = p(u_i,u_j)\log\frac{p(u_i,u_j)}{p(u_j)p(u_j)}, \qquad (15.3)$$

где u_i – SWU с номером i, полученная в предыдущих циклах итераций.

На каждой итерации n пар SWU (n устанавливается эмпирически) с наибольшими взвешенными взаимными информациями объединяется. Конечное количество SWU зависит от числа итераций. Попутно можно получить парсинг всего словаря в терминах, полученных SWU. Эксперименты с английским языком продемонстрировали, что после 200 итераций из получившихся 1977 SWU две трети представляли обычные слоги, а средняя длина SWU составляла 3.2 фонемы.

Метод показал существенное улучшение по сравнению с фонемными N-граммами: для уровня определения OOV-слов в 70% ошибка ложной тревоги уменьшилась с 8.7% до 3.2%.

Развитием описанных методов является мультиклассовая OOV-модель. Идея модели заключается в том, что все слова можно разбить на грамматические классы (существительные, глаголы, прилагательные, наречия, а также имена). Слова, принадлежащие к одному классу, могут иметь сходства в фонотактической структуре и выступать сходным образом в модели языка, что позволяет получить более точную модель OOV-слова в соответствии с его местом во фразе.

15.3. Использование нескольких систем распознавания

Новым подходом в детектировании OOV-слов является использование доверительных оценок, полученных распознающими системами с различными по степени ограничениями. Например, в работах [49, 50] сравнивались апостериорные вероятности, полученные от системы с сильными ограничениями (система распознавания слитной речи с большим словарём и моделью языка – LVCSR – Large Vocabulary Continuous Speech Recognition system) и системы со слабыми ограничениями (система распознавания фонемных цепочек). Метод основан на естественном предположении, что, если отсутствующее в словаре слово будет ошибочно распознано как словарное, то цепочка распознанных фонем будет содержать фрагменты, плохо согласующиеся с распознанным словом. Использование фонемного распознавания стало возможным в последние годы на основе применения долговременных признаков (см. 10.1) и нейронных

сетей для оценки вектора апостериорных вероятностей фонем, соответствующих фрагменту речи.

В результате работы двух систем распознавания для каждого момента времени *t* будут получены два вектора, представляющие вероятности фонем:

$$\begin{aligned} P_t &= (p(1|t),\ldots,p(i|t),\ldots,p(N_p|t))^T, \\ Q_t &= (q(1|t),\ldots,q(i|t),\ldots,q(N_p|t))^T \end{aligned}, \qquad (15.4)$$

где N_p – количество фонем, $p(i|t)$ и $q(i|t)$ – апостериорные вероятности для фонемы с номером *i* в момент времени *t* для двух систем.

Эти данные используются для получения локальных (на одном окне) доверительных оценок, на основании которых можно выносить суждение о наличии или отсутствии внесловарного слова. Отметим, что векторы вероятностей, полученные системой с сильными ограничениями, использующими словарь и модель языка, в отличие от модели со слабыми ограничениями, содержат большое количество компонент, близких к нулю. Это затрудняет применение стандартных метрик, использующих логарифмы, например, метрики Кульбака-Лейбрера. Наилучший результат даёт подход, основанный на нейронных сетях [49]. Усредняя на слове локальные доверительные оценки, получим интегральную доверительную оценку для данного слова, при этом примем во внимание, что, если OOV-слово отличается от IV-слова только одно фонемой, такое усреднение не имеет смысла.

Следует отметить, что, несмотря на некоторые успехи в определении внесловарных слов, задача остаётся до конца не решённой – существуют слова, которые являются объединением словарных слов (female – fee male), определить которые может только система с мощной моделью языка, а то и с элементами понимания; возможны оговорки, неправильные произнесения незнакомых для диктора терминов и много другого брака в акустическом сигнале. По-видимому, говорить о «полном решении» задачи определения OOV-слов можно в том же смысле, что и об «окончательном решении» задачи автоматического распознавания речи.

16. АУДИОВИЗУАЛЬНОЕ РАСПОЗНАВАНИЕ РЕЧИ

Во многих условиях функционирования (низкое качество звукового сигнала, присутствие сильного внешнего шума или посторонних разговоров и т.д.) стандартные системы автоматического распознавания речи не могут обеспечивать приемлемое качество работы даже при применении различных методов фильтрации и шумоподавления. Для того чтобы повысить качество и робастность работы автоматических систем применяются также способы распознавания визуальной информации о речи на базе технологий машинного зрения (так называемое «чтение речи по губам»), создавая системы аудиовизуального (бимодального) распознавания речи. Очевидно, что речь передается не только в виде звуковой волны, она поступает от человека одновременно по нескольким информационным каналам (модальностям), в том числе по звуковому и визуальному. Так, например некоторые реализации фонем (звуков речи) очень легко спутать на слух (например, /м/ и /н/), но легко отличить визуально (/м/ производится с закрытым ртом, а /н/ – с открытым). При объединении потоков информации от аудио- и видеораспознавателей результат совместной обработки может превышать точность распознавания по каждой из модальностей. При восприятии речи человеком известен также эффект МакГурка (McGurk) [80], когда правильный элемент возникает только при объединении звуковой информации и визуальной информации, получаемой слушателем от артикуляции губ диктора.

Как известно, речь производится человеком путем взаимосвязанных действий нескольких групп анатомических органов человека (грудная клетка, легкие, трахея, голосовые связки, гортанная трубка, полость глотки, нёбная занавеска, полость рта, полость носа, язык, губы) [56]. В ходе комплексного процесса понимания речи органы слуха человека воспринимают звуки, в то время как глаза видят движения лица, языка и губ и вся эта информация объединяется в мозгу человека в единое представление смысла высказывания. Слабослышащие и пожилые люди, а также неносители языка больше опираются на визуальную информацию, выражаемую движениями губ и лицевыми органами, чем на звуковую. Визуальные сигналы очень важны для лучшего понимания произносимой речи, так, глядя в лицо собеседнику, легче понимать его речь, особенно если речь иностранная. Сигналы от визуальных и слуховых каналов дублируют и дополняют друг друга, что помогает правильно воспринимать речь во многих сложных ситуациях, например, при воздействии динамических акустических шумов, или когда одновременно говорят несколько человек.

16.1. Способы объединения аудио- и видеомодальностей речи

Существуют два основных подхода к объединению звуковой и визуальной информации (information fusion) при бимодальном распознавании речи [81]:

1) Первый способ называют «ранним» объединением (early fusion) – рисунок 16.1. В данном подходе независимо вычисляется параметрическое представление звукового и визуального сигналов, а затем, с учетом достаточно

высокой степени синхронности этих модальностей, формируется единый вектор признаков (супервектор) для каждого сегмента сигнала. На этапе классификации (распознавания) речи используются методы, использующие Скрытые Марковские Модели (СММ) или Искусственные Нейронные Сети, при этом создаются общие модели для акустических единиц речи (фонем) и визуальных единиц речи (визем – изображений формы губ при произнесении различных фонем).

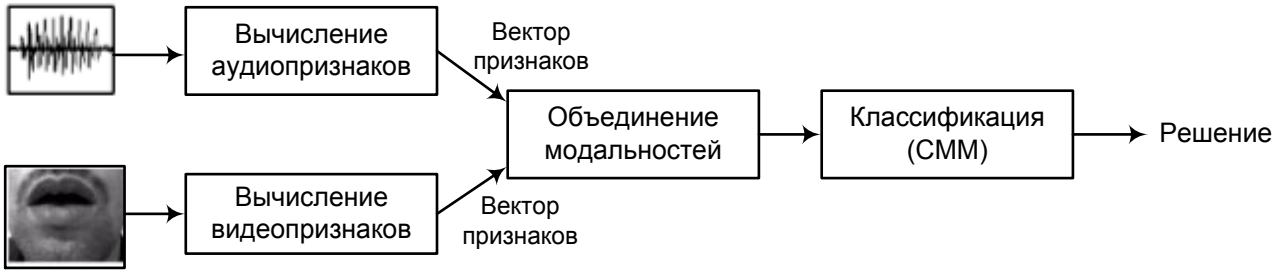


Рис. 16.1. Способ «раннего» объединения звуковой и визуальной модальностей.

2) Второй способ объединения информации осуществляет «позднее» объединение модальностей (late fusion) – рисунок 16.2. Способ поздней интеграции использует независимые друг от друга СММ для вероятностного моделирования звуковых и визуальных сигналов речи. Объединение модальностей возможно как на уровне состояний вероятностных моделей, так и на уровне потоков фонем/визем или даже гипотез распознавания фраз.

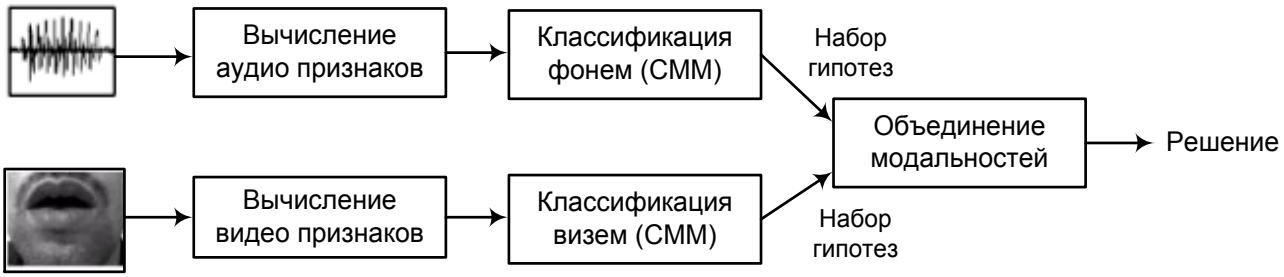


Рис. 16.2. Способ «позднего» объединения звуковой и визуальной модальностей речи.

Существуют также и различные гибридные способы объединения аудиовизуальной речевой информации. При этом, в любом способе объединения бимодальной информации окончательное решение о распознанном сообщении принимается с учетом весовых коэффициентов информативности каждой модальности, которые изменяются в зависимости от окружающих условий (уровень шума, освещения, и т.д.). Так, например, информативность звуковой информации будет невысока в производственных помещениях, в аэропортах и вокзалах, в условиях сильного дождя и т.д. Визуальная же информация будет практически бесполезна при слабом освещении, ночью и т.д. Объединение же

модальностей позволяет получить правильный результат практически в любых условиях эксплуатации, повышая робастность к шумам и точность работы системы распознавания речи.

Реализация метода синхронизации и интеграции речевых модальностей является одной из основных задач системы аудиовизуального распознавания речи на любом языке. Суть проблемы состоит в естественном рассогласовании двух основных речевых модальностей, т.е. в естественной человеческой речи потоки соответствующих фонем и визем (которые являются результатом артикуляции) не являются полностью синхронными по времени, хотя и значительно перекрываются [82]. Так, например, при произнесении звука /у/ мы сначала складываем губы “в трубочку”, а затем производим нужный звук или же для произнесения звука /м/ мы сначала должны сомкнуть губы. Такой феномен, вызван динамикой процесса речеобразования (ограниченной скоростью движения органов артикуляции, которые и формируют различные звуки) и эффектом коартикуляции, который является процессом взаимного влияния (взаимопроникновения) определенных речевых единиц на соседние элементы речи. Необходимо отметить, что коартикуляция по-разному проявляется на акустическом и визуальном компонентах речи, что и вызывает асинхронность между ними. Правильная синхронизация модальностей речи имеет важное значение и при восприятии речи человеком (в том числе синтезированной речи), так как она напрямую влияет как на разборчивость, так и естественность речевого высказывания. Исследования показывают, что для различных языков и культур степень синхронности потоков фонем и визем в процессе речеобразования различна. Так, например, для японского языка движения губ и звуковой поток практически синхронны, поэтому способ раннего объединения многомодальной информации показывает наилучшие результаты [83]. Английскому же языку (особенно американскому варианту) сопутствует достаточно богатая артикуляция губ (даже гиперартикуляция), что вызывает определенные временные расхождения между потоками фонем и визем (до нескольких сотен миллисекунд); поэтому модели распознавания с поздним объединением информации в этом случае предпочтительны.

В разных языках эксперты выделают разное количество визем (также как и фонем), в английском их 12-14, в русском 10-12 в зависимости от диктора. В таблице 16.1 показаны базовые классы визем русской речи и соответствие фонемам. В данном случае мы считаем, что в русской речи существует 47 фонем, включая ударные и безударные (редуцированные) варианты гласных; согласные звуки соответствуют фонетическому алфавиту SAMPA. В данной таблице представлены 10 классов визем, включая нейтральное положение губ.

Таблица 16.1.

Классы визем русской речи и их соответствие фонемам русской речи.

Класс виземы	Тип виземы/фонемы	Соответствующие фонемы русской речи
V1	Неогубленные гласные (широкое открытие рта)	/а!/, /а/, /э!/, /э/
V2	Неогубленные гласные (остальные)	/и/, /и!/, /ы/, /ы!/
V3	Огубленные гласные звуки	/о!/, /у/, /у!/
V4	Губные согласные	/б/, /б’/, /п/, /п’/, /м/, /м’/
V5	Губно-зубные согласные	/ф/, /ф’/, /в/, /в’/
V6	Альвеолярные фрикативные согласные	/ш/, /ж/, /ч/, /щ/
V7	Альвеолярные сонорные согласные	/л/, /л’/, /р/, /р’/
V8	Зубные согласные	/д/, /д’/, /т/, /т’/, /н/, /н’/, /с/, /с’/, /з/, /з’/, /ц/
V9	Заднеязычные согласные	/г/, /г’/, /к/, /к’/, /х/, /х’/, /й/
V10	Пауза (нейтральное положение губ)	тишина (пауза)

На рисунке 16.3 показана общая архитектура системы аудиовизуального распознавания речи. В системе используются независимые друг от друга цифровая видеокамера и микрофон, параллельно вычисляются признаки аудио- и видеосигналов, объединение модальностей может производиться на раннем либо позднем уровне.

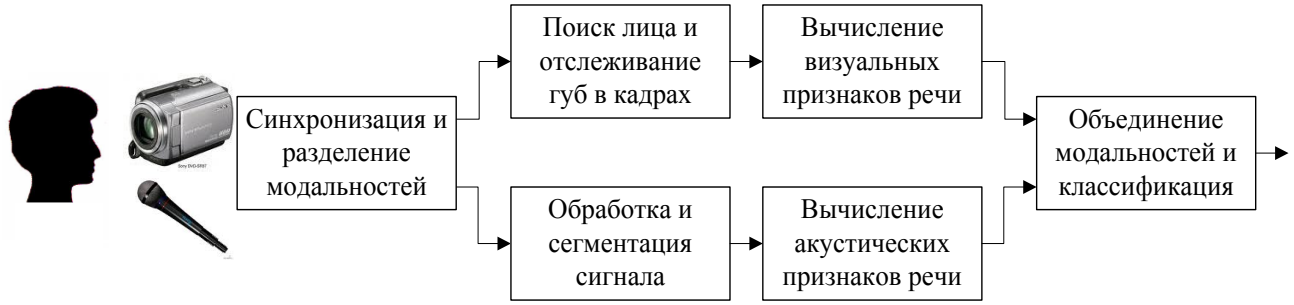


Рис. 16.3. Архитектура системы аудиовизуального распознавания речи.

Для вычисления информативных признаков из аудиосигнала могут использоваться различные методы спектральной обработки (например, MFCC признаки), которые были описаны в предыдущих разделах. Для вычисления признаков движений губ по видеосигналу могут использоваться 2 различных подхода:

1) Пиксельные признаки: нахождение прямоугольной графической области (region of interest) рта диктора и извлечение информативных признаков (например, методом анализа главных компонент PCA).

2) Геометрические визуальные признаки: нахождение и описание контура (формы) губ диктора при говорении (например, используя цветовую фильтрацию изображения); такие признаки описывают геометрические параметры губ диктора: ширина рта, толщина верхней и нижней губ, видимость языка и зубов и т.д.

16.2 Методы аудиовизуального моделирования и распознавания речи

Наиболее распространены два основных метода моделирования и распознавания аудиовизуальной речи, которые основаны на следующих модификациях скрытых марковских моделей:

1) Многопоточные скрытые марковские модели (МПСММ, Multi-stream HMM), в которых независимо используются несколько типов признаков в единой СММ. Эти модели реализуют ранний подход к объединению модальностей (на уровне признакового описания), такие модели являются синхронными относительно аудио- и видеопризнаков речи.

2) Сдвоенные скрытые марковские модели (ССММ, Coupled HMM), в которых используются параллельные асинхронные СММ модели, а объединение информации происходит на уровне состояний ССММ моделей, реализуя поздний подход к объединению информации.

В данных типах моделей звуковые и визуальные речевые признаки разносятся по двум разным потокам, но объединение происходит различными способами на разных уровнях скрытых марковских моделей. На рисунке 16.4. показан пример топологии МПСММ одной аудиовизуальной единицы речи [84], здесь видно, что состояния в моделях являются общими для двух потоков информации, а вектора признаков аудио- и видеосигналов вычисляются независимо и объединяются с учетом весовых коэффициентов каждой модальности. Недостатком таких моделей является то, что они не могут обрабатывать возможную асинхронность аудио- и видеомодальностей речи.

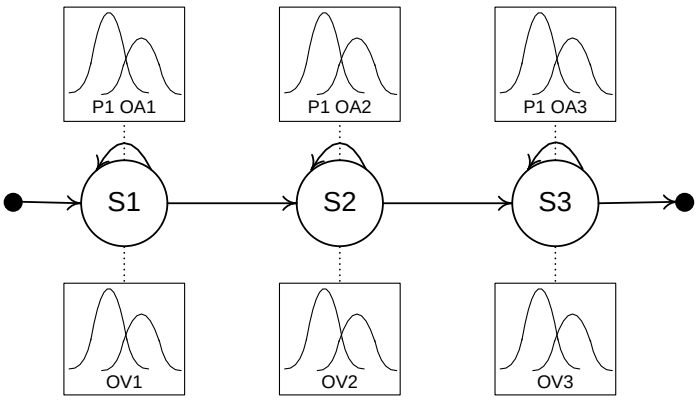


Рис. 16.4. Топология многопоточной СММ аудиовизуальной единицы речи.

Для учета естественной для речеобразования временной асинхронности потоков соответствующих акустических и визуальных признаков речи были

разработаны сдвоенные скрытые марковские модели (CCMM, Coupled Hidden Markov Model) [85]. На рисунке 16.5 показана топология модели аудиовизуальной единицы речи (пара фонема/визема) с несколькими состояниями для каждого потока векторов признаков. Кругами обозначены состояния CCMM, являющиеся скрытыми для наблюдения, а квадратами – смеси нормальных распределений векторов наблюдений в состояниях. Сдвоенная скрытая марковская модель представляет собой набор параллельных CMM, по одной на каждый информационный поток (модальность), состояния модели в некоторый момент времени *t* для каждой CMM зависят от скрытых состояний в момент времени *t-1* всех параллельных CMM. Таким образом, общее состояние CCMM определяется совокупностью состояний двух параллельных CMM. Преимущество такой топологии состоит в том, что она позволяет нескольким потокам векторов признаков независимо переходить по состояниям модели, что дает возможность моделировать допустимые временные расхождения в аудио- и видеоданных. В топологии CCMM аудиовизуальных единиц речи применяются по три состояния на каждый параллельный поток векторов признаков, при этом считается, что первые состояния соответствуют динамическому переходу от предыдущей речевой единицы, третьи – переходу к последующей единице, а вторые состояния объединенной модели (самые длительные) соответствуют стационарному (центральному) участку речевой единицы.

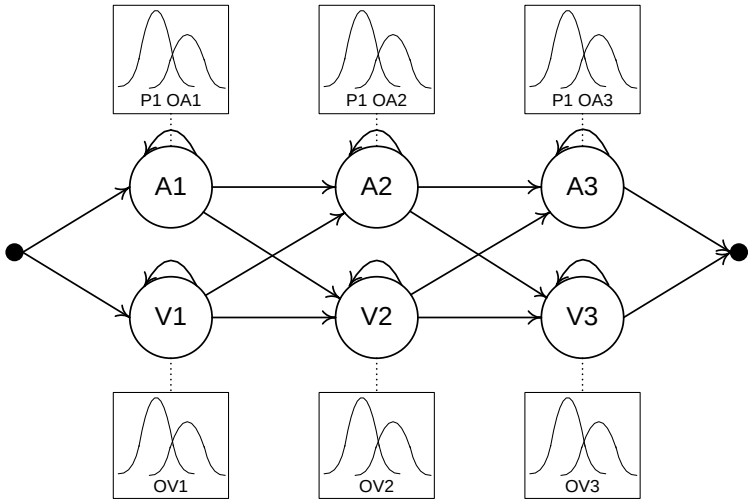


Рис. 16.5. Топология сдвоенной CMM аудиовизуальной единицы речи.

Для определения сдвоенной скрытой марковской модели $\lambda =< L, D, B, \gamma >$ некоторой аудиовизуальной единицы речи необходимо задать следующие параметры:

1) Количество скрытых состояний модели – L.

2) Распределение (матрица) вероятностей переходов между состояниями - $D = \{d_{ij}\}$, $1 \le i \le L$, $1 \le j \le L$.

3) Распределение вероятностей появления символов наблюдения (векторов признаков аудиовизуальной речи) в состояниях - $B = \{b_j(O)\}$. Обычно применяются смеси нормальных (гауссовских) распределений вероятностей:

$$b_j(O) = \sum_{m=1}^{M} c_{jm} N(O, \mu_{jm}, \sigma_{jm}), \sum_{m=1}^{M} c_{jm} = 1, \quad (16.1)$$

где O - моделируемый вектор параметров (аудио или видеосигнала), C_{jm} - весовой коэффициент m-й компоненты в состоянии j, N - плотность вероятности (гауссовское распределение) со средним значение (математическое ожидание) μ_{jm} и среднеквадратическое отклонение (дисперсия) σ_{jm} для m-й компоненты смеси в состоянии j, M – количество смесей нормальных распределений в модели.

4) Веса информативности (значимость) $\gamma = \{\gamma^A, \gamma^V\}$ речевых модальностей (аудио- и видеопотоков), которые могут настраиваться в ходе обучения моделей или адаптации к окружающим условиям и каналу передачи речи.. Причем, их сумма является константой:

$$\gamma^A + \gamma^V = 2 \quad (16.2)$$

В русской речи фонетисты выделяют несколько десятков различных фонем (от 40 до 50), поэтому и ССММ в бимодальной системе распознавания речи насчитывается столько же. Различимых единиц видимой русской речи (визем) намного меньше (около 10). Поэтому, как правило, в системе распознавания применяется связывание (tying) распределений векторов наблюдений визуальных компонент в состояниях разных ССММ. На рисунке 16.6 показаны связи параметров (распределения векторов наблюдений) визуальных моделей в рамках одного класса виземы при наличии нескольких акустических моделей (например, одна визема для двух фонем /б/ и /м/). Таким образом, общее количество ССММ в системе равняется числу распознаваемых фонем, но для ряда моделей их параметры являются общими, что упрощает и улучшает процесс обучения моделей в условиях ограниченных обучающих данных.

Кроме того, в работе [86] был предложен весьма простой способ преобразования топологии сдвоенной СММ в эквивалентную лево-правую двухпоточную СММ модель (см. рисунок 16.7), которая сохраняет все свойства первой. Результирующая СММ содержит все комбинации параллельных состояний исходной ССММ. В ССММ оба потока независимы и распределения векторов наблюдений в состояниях вычисляются отдельно друг от друга, в двухпоточной СММ два распределения векторов наблюдений (для аудио- и видеосигналов) ассоциированы с каждым состоянием. В топологии ССММ используется по 3 скрытых состояния на каждый параллельный поток, соответственно в двухпоточной СММ будет 9 состояний (все их комбинации), поэтому, для того чтобы избежать утроения количества распределений векторов наблюдений в состояниях, используется связывание соответствующих распределений векторов наблюдений согласно рисунку 16.7.

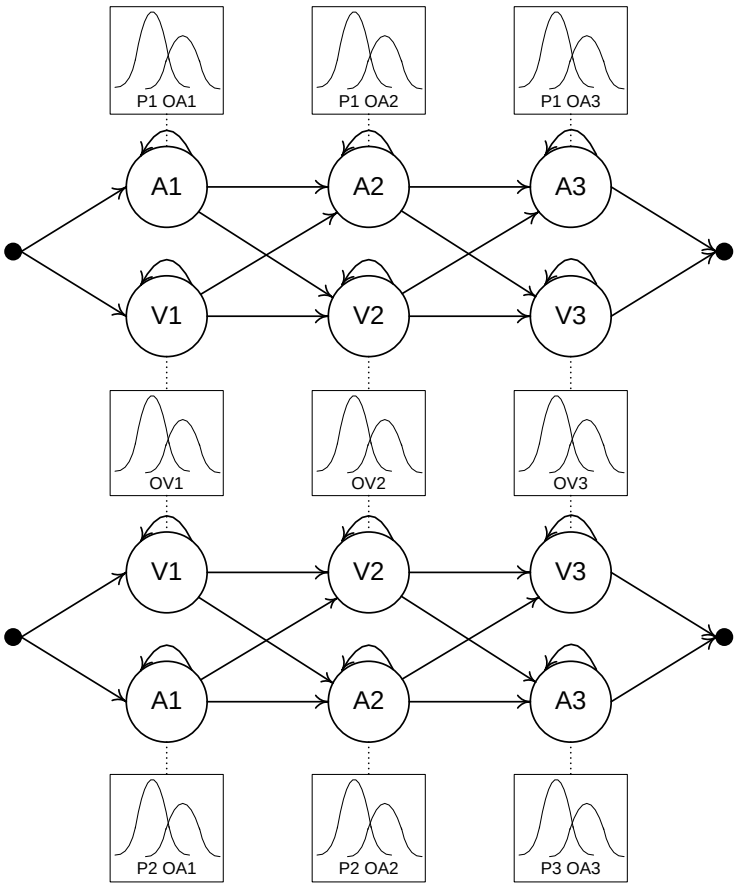


Рис 16.6. Связывание двух CMMM для двух аудиовизуальных единиц русской речи из одного виземного класса.

Увеличение общего количества состояний модели аудиовизуальной единицы речи с 6 до 9 требует несколько большей оперативной памяти при программной реализации системы, однако на скорости декодирования такое преобразования не сказывается из-за большей простоты алгоритма декодирования речи (алгоритм Витерби, был описан в предыдущих разделах). Параметры двухпоточных CMM (матрица вероятностей переходов из состояния в состояние и распределения векторов признаков аудио- и видеоданных в скрытых состояниях) аудиовизуальных единиц речи вычисляются, используя модифицированный алгоритм Баума-Уэлча (EM-алгоритм) [87], добиваясь максимальной оценки правдоподобия модели на обучающей выборке мультимедийных речевых записей.

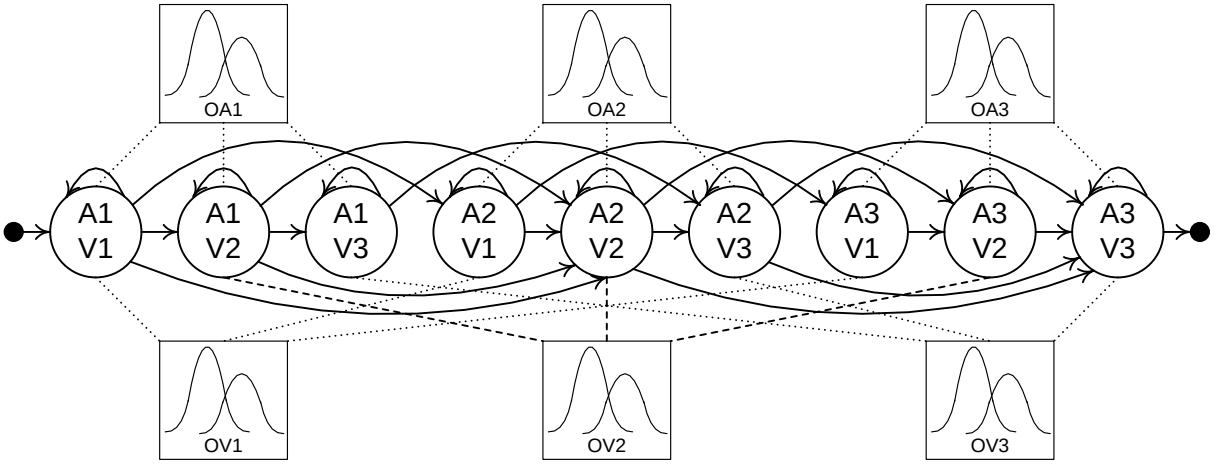


Рис 16.7. Отображение сдвоенной СММ в эквивалентную двухпоточную СММ.

Для декодирования (распознавания) слитной речи, подаваемой на вход системы из аудиовизуального файла или с устройств, применяется модифицированный метод передачи маркеров (token-passing method) [84], основанный на оптимизационном алгоритме Витерби для многопоточных СММ, который определяет вероятность порождения символов наблюдений (последовательностей векторов признаков) данной моделью и последовательность пройденных при этом скрытых состояний модели. Суть данного метода состоит в следующем: для моделирования потенциально возможных фраз слитной речи строится единая вероятностная модель (граф) со всевозможными вариантами переходов между СММ минимальных единиц речи (ограниченными словарем распознавания) и между СММ слов (ограниченными конечной грамматикой или вероятностной моделью языка), затем методом динамического программирования (алгоритм Витерби) производится нахождение оптимальной по критерию максимального правдоподобия последовательности/пути скрытых состояний (несущих информацию о речевых единицах) модели для порождения обрабатываемой последовательности наблюдений. В результате декодирования речевого сигнала автоматической системой выдается одна или несколько наилучших гипотез (N-best list) распознавания произнесенной диктором фразы.

Многие экспериментальные результаты мировых исследований по созданию систем бимодального распознавания речи, которые используют аудиовизуальную информацию, показывают, что они работают лучше (т.е. обеспечивают более высокую точность распознавания речи и лучшую робастность системы к аудиошумам), чем одномодальные системы, использующие только звуковую информацию [88-92].

ЛИТЕРАТУРА

1. Фланаган Дж. Анализ, синтез и восприятие речи. «Связь». Москва, 1968.
2. MIT Lectures 2003. http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-345-automatic-speech-recognition-spring-2003/download-course-materials/
3. Фант. Г. Акустическая теория речеобразования. «Наука». Москва, 1964.
4. Picone J. Fundamentals of speech recognition: a short course.1996. http://speech.tifr.res.in/tutorials/fundamentalOfASR_picone96.pdf
5. Алдошина И. Основы психоакустики. http://giga.kadva.ru/files/edu/AldoshinaPsychoacoustics.pdf
6. Слуховая система. серия "Основы современной физиологии". «Наука». Ленинград, 1990.
7. Seneff S. “Pitch and Spectral Analysis of Speech Based on an Auditory Synchrony Model”, Technical Report 504**,** January 1985
8. Hermansky H. (1997): “Should recognizers have ears?”, In RSR-1997, 1-10.
9. Маркел Дж.Д., Грей А.Х., Линейное предсказание речи, Москва, «Связь». 1980.
10. Hermansky H., Morgan N., "RASTA Processing of Speech", in IEEE Transaction on Speech and Audio Processing, Vol. 2, No. 4, pp. 587-589, October 1994.
11. Карпов А.А., Кипяткова И.С., Методология оценивания работы систем автоматического распознавания речи // Известия вузов. Приборостроение, Т. 55, № 11, 2012, С. 38-43.
12. Левенштейн В.И., Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий Наук СССР, 1965, 163.4:845-848.
13. Khokhlov Y., Tomashenko N., “Speech Recognition Performance Evaluation for LVCSR System”, In Proc. 14th International Conference “Speech and Computer” SPECOM-2011, Kazan, Russia, 2011, pp. 129-135.
14. Kurimo M., Creutz M., Varjokallio M., Arsoy E., Saraclar M., Unsupervised segmentation of words into morphemes - Morpho challenge 2005 Application to automatic speech recognition. In Proc. INTERSPEECH 2006, Pittsburgh, USA, 2006, pp. 1021-1024.
15. Schlippe T., Ochs S., Schultz T., Grapheme-to-Phoneme Model Generation for Indo-European Languages. In Proc. ICASSP-2012, Kyoto, Japan, 2012.
16. Huang C., Chang E., Zhou J., Lee K. Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition. In Proc. INTERSPEECH 2000, Beijing, China, 2000, pp. 818-821.
17. Ablimit M., Neubig G., Mimura M., Mori S., Kawahara T., Hamdulla A. Uyghur Morpheme-based Language Models and ASR. In Proc. 10th IEEE International Conference on Signal Processing ICSP-2010, Beijing, China, 2010, pp. 581-584.

18. Karpov A., Kipyatkova I., Ronzhin A. Very Large Vocabulary ASR for Spoken Russian with Syntactic and Morphemic Analysis. In Proc. INTERSPEECH 2011, Florence, Italy, 2011, pp. 3161-3164.
19. Nanjo H., Kawahara T. A New ASR Evaluation Measure and Minimum Bayes-Risk Decoding for Open-domain Speech Understanding. In Proc. ICASSP-2005, Philadelphia, USA, 2005, pp. 1053-1056.
20. The US NIST 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan, http://www.itl.nist.gov/iad/mig/tests/rt/2009/
21. Morris A.C., Maier V., Green P. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition, In Proc. INTERSPEECH 2004, Jeju Island, Korea, 2004, pp. 2765-2768.
22. Tran B.-H., Seide F., Steinbiss T. A word graph based N-best search in continuous speech recognition. In Proc. ICSLP-96, Philadelphia, USA, 1996, pp. 2127-2130.
23. Vilar J.M. Efficient computation of confidence intervals for word error rates, In Proc. ICASSP-2008, Las Vegas, USA, 2008, pp. 5101-5104.
24. Hansen J.H., Clements M.A. "Spectral Slope based Distortion Measures for All-Pole Models of Speech", Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, pp.757-760, 1986.
25. Mansour D., Juang B.H. "A Family of Distourtion Measures based upon Projection Operation for Robust Speech Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing., Vol.37, No.11, pp. 1659-1671, November 1989.
26. MIT lectures "Automatic Speech Recognition", Lecture 9, **"**Dynamic Time Warping & Search", http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-345-automatic-speech-recognition-spring-2003/lecture-notes/lecture9.pdf.
27. Рабинер Л. Скрытые марковские модели и их применение в избранных приложениях при распознавании речи: Обзор. ТИИЭР, 1989, т. 77, № 2, с. 86-120. Rabiner L.R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, Vol. 77, No. 2, February 1989, pp. 257-286. http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf
28. Burshtein D., "Robust Parametric Modeling of Durations in Hidden Markov Models", IEEE Transactions on Speech and Audio Processing, Vol. , No. 3, May 1996, pp. 240-242.
29. Odell J.J., "The Use of Context in Large Vocabulary Speech Recognition", Dissertation, 1995. http://mi.eng.cam.ac.uk/reports/svr-ftp/auto-pdf/odell_thesis.pdpdf/odell_thesis.pdf
30. Huang X., Acero A., Hon H.-W. Spoken language processing. 2001.
31. Legetter C.J., "Improved Acoustic Modelling for HMMs Using Linear Transformations", Ph.D. thesis, Cambridge University, 1995. ftp://svr-ftp.eng.cam.ac.uk/pub/reports/auto-pdf/leggetter_thesis.pdf
32. Jeff A. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models",

International Computer Science Institute Berkeley CA, 94704 and Computer Science Division Department of Electrical Engineering and Computer Science, U.C. Berkeley, TR-97-021, April 1998.

33. Nguen P., “Fast Speaker Adaptation”, Rapport de these professionnelle, Institut Eurecom, June 18, 1998.
34. Kuhn R., Junqua J.-C., Nguen P., Niedzielski N., “Rapid Speaker Adaptation in Eigenvoice Space”, IEEE Transactions on Speech and Audio Processing, Vol. 8, No. 6, November 2000.
35. Jonson K., “Speaker Normalization in Speech Perception”, Ohio State University, 2005, pp. 1-45.
36. Panchapagesan S., “Frequency Warping by Linear Transformation of Standard MFCC,” in Interspeech 2006, Pittsburgh, USA, 2006.
37. Molau S., Kanthak S., Ney H., “Efficient Vocal Tract Normalization in Automatic Speech Recognition”. Konf. Elektron. Sprachsignalverarbeitung, Cottbus, pp. 209-216, Sep. 2000.
38. Hain T., Woodland P., Niesler T., Whittacker E., “The 1998 HTK System For Transcription of Conversational Telephone Speech”, in: Proceedings International Conference on Acoustics, Speech and Signal Processing, Mar 1999, pp. 57 – 60, vol.1.
39. Hermansky H., Sharma S., "Temporal patterns (traps) in asr of noisy speech", in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP), Phoenix, Arizona, USA, Mar. 1999. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.135.5036
40. Hermansky H., Jain P., "Beyond a single critical-band in trap based asr", in Proc. Eurospeech, Geneva, Switzerland, Sep. 2003.
41. Schwarz P., "Phoneme recognition based on long temporal context", Ph.D. thesis, Brno University of Technology, 2008. http://www.fit.vutbr.cz/~schwarzp/publi/thesis.pdf
42. Oparin I., Talanov A. “Stem-Based Approach to Pronunciation Vocabulary Construction and Language Modeling of Russian”, Proc. of the 10th International conference on Speech and Computer, SPECOM-2005. Patras, Greece, 2005. pp. 575-578.
43. Jelinek F., Chelba C. Recognition Performance of a Structured Language Model. Eurospeech 1999. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.2129
44. Mohri M., Pereira F., Riley M., “SPEECH RECOGNITION WITH WEIGHTED FINITE-STATE TRANSDUCERS”, Springer Handbook on Speech Processing and Speech Communication.
45. Kinderman R., Snell J.L., “Markov Random Fields and Their Applications”, American Mathematical Society, 1980. http://www.cmap.polytechnique.fr/~rama/ehess/mrfbook.pdf
46. Abdel-Haleem Y.H., “Conditional Random Fields for Continuous Speech Recognition”, 2006. http://homepages.inf.ed.ac.uk/srenals/yasser-thesis.pdf

47. Gunawardana A., Mahajan M., Acero A., Platt J. C., “Hidden conditional random fields for phone classification”// In Proc. INTERSPEECH 2005, pp. 1117–1120.
48. Parada C., Dredze M., Sethy A., Rastow A., “ Learning Sub-Word Units for Open Vocabulary Speech Recognition”, Proc. Of the 49th Annual Meeting of the Association for Computational Linguistics, June 19-24, 2011, pp. 712-721.
49. Burget L., Schwarz P., et al., “COMBINATION OF STRONGLY AND WEAKLY CONSTRAINED RECOGNIZERS FOR RELIABLE DETECTION OF OOVS”, IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008, pp. 4081-4084.
50. Hannemann M., “Combinations of Confidence Measures for the Detection of Out-of-Vocabulary Segments in Large Vocabulary Continuous Speech Using Differently Constrained Recognizers”, Otto-von-Guericke-Universitat Magdeburg, 21. April 2008.
51. Bourlard H., Wellekens C.J., “Links between Markov models and multilayer perceptrons”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12 , No. 12, 1990, pp. 1167-1178.
52. Bourlard H., Hermansky H., Morgan N., “Towards increasing speech recognition error rates”, Speech Communication, Vol. 18, 1996, p.p. 205–231.
53. Hornik K., Stinchcombe M., White H., “Multilayer feedforward networks are universal approximators”, Neural Netw. Vol. 2(5), 1989, pp. 359–366.
54. Hinton G., Deng L., Yu D., Dahl G., Mohamed A., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T., Kingsbury B., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”, IEEE Signal Process. Mag., Vol. 29, No. 6, Nov. 2012, pp. 82–97.
55. Dong Yu, Li Deng, “Automatic Speech Recognition. A Deep Learning Approach”, Springer-Verlag, London. 2015, 321 p.
56. Чистович Л.А. и др., «Руководство по физиологии. Физиология речи. Восприятие речи человеком», «Наука», Ленинград, 1976.
57. Hermansky H., Ellis D., Sharma S., “Tandem connectionist feature extraction for conventional HMM systems”, Proc. ICASSP-2000, Istanbul. 2000. V. 3. pp. 1635–1638.
58. Deng, L., Chen, J., “Sequence classification using high-level features extracted from deep neural networks.” In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 6894-6898.
59. Hochreiter S., Schmidhuber J., “Long short-term memory.” Neural Computation, V. 9(8), 1997, pp. 1735–1780.
60. Pascanu R., Mikolov T., Bengio Y., “On the difficulty of training recurrent neural networks”, Cornell University Library, arXiv:1211.5063 [cs.LG], 2013.
61. Graves A., Fernández S., Schmidhuber J., “Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition”, Chapter: **“**Artificial Neural Networks: Formal Models and Their Applications” – ICANN 2005, pp 799-804.

62. Wollmer M., Eyben F., Schuller B., Rigoll G., “Recognition of Spontaneous Conversational Speech using Long Short-Term Memory Phoneme Predictions”, In: INTERSPEECH 2010, pp. 1946-1949.
63. Graves A., Mohamed A., Hinton G., “Speech recognition with deep recurrent neural networks”, Cornell University Library, arXiv:1303.5778 [cs.NE], 2013.
64. Graves A., Jaitly N., “Towards End-to-End Speech Recognition with Recurrent Neural Networks”, Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP V. 32.
65. Sak H., Senior A., Beaufays F., “Long short-term memory recurrent neural network architectures for large scale acoustic modeling”, INTERSPEECH 2014, pp. 338-342.
66. Sak H., Vinyals O., Heigold G., Senior A., McDermott E., Monga R., Mao M., “Sequence discriminative distributed training of long short-term memory recurrent neural networks”, In: INTERSPEECH 2014.
67. Triefenbach F., Jalalvand A., Schrauwen B., Martens J.-P., “Phoneme Recognition with Large Hierarchical Reservoirs”, Conf: Advances in Neural Information Processing Systems 23, 2010, pp. 2307-2315.
68. Triefenbach F., Demuynck K., Martens J.-P., “Improving large vocabulary continuous speech recognition by combining gmm-based and reservoir-based acoustic modeling”, Spoken Language Technology Workshop (SLT), 2012 IEEE, pp. 107-112.
69. Triefenbach F., Demuynck K., Martens J.-P., “Large vocabulary continuous speech recognition with reservoir-based acoustic models”, IEEE Signal Processing Letters, Vol. 21, No. 3, March 2014, pp. 311-315.
70. Jalalvand A., Demuynck K., De Neve W., Van de Walle R., Martens J.-P., “Design of Reservoir Computing Systems for Noise-Robust Speech and Handwriting Recognition”, Multimedia Lab, ELIS, Ghent University. Ghent 9000, Belgium, 2015.
71. Ronzhin A., Karpov A., “Implementation of morphemic analysis for Russian speech recognition”, In Proc. 9th International Conference on Speech and Computer SPECOM-2004, St. Petersburg, Russia, 2004, pp. 291-296.
72. Creutz M., Hirsimaki T., Kurimo M., Puurula A., Pylkkonen J., Siivola V., Varjokallio M., Arisoy E., Saraclar M., Stolcke A., “Morph-based speech recognition and modeling of out-of-vocabulary words across languages”, ACM Transactions on Speech and Language Processing, 5(1), 2007.
73. Sak H., Saraclar M., Güngör T., “Morphology-based and sub-word language modeling for Turkish speech recognition”. In Proc. ICASSP-2010, pp. 5402-5405.
74. Tarjan B., Mihajlik P., “On Morph-Based LVCSR Improvements”, In Proc. 2nd International Workshop on Spoken Languages Technologies for Under-resourced Languages SLTU-2010, Malaysia, 2010, pp. 10-16.
75. Chelba C., Jelinek F., “Structured language model” // Computer Speech and Language, 2000. Vol. 10. pp. 283-332.

76. Sidorov G., Velasquez F., Stamatatos E., Gelbukh A., Chanona-Hernández L., “Syntactic Dependency-based N-grams as Classification Features”, Springer LNAI 7630, Mexico, 2012. pp. 1-11.
77. Szarvas M., Furui S., “Finite-state transducer based modeling of morphosyntax with applications to Hungarian LVCSR” // Proc. ICASSP’2003, Hong Kong, China, 2003. pp. 368–371.
78. Холоденко А.Б., “О построении статистических языковых моделей для систем распознавания русской речи” // Интеллектуальные системы, 2002. Т.6. Вып. 1-4. С. 381-394.
79. Karpov A., Markov K., Kipyatkova I., Vazhenina D., Ronzhin A., “Large vocabulary Russian speech recognition using syntactico-statistical language modeling” // Speech Communication. 2014, Vol. 56, pp. 213-228.
80. McGurk H., MacDonald J., “Hearing Lips and Seeing Voices // Nature, 264, 1976, pp. 746-748.
81. Карпов А.А., “Реализация автоматической системы многомодального распознавания речи по аудио- и видеоинформации” // Автоматика и телемеханика. 2014, Т. 75, № 12, С. 125-138.
82. Кипяткова И.С., Ронжин А.Л., Карпов А.А., “Автоматическая обработка разговорной русской речи”. – СПб.: ГУАП, 2013. – 314 с.
83. Sekiyama K., “Differences in auditory-visual speech perception between Japanese and America: McGurk effect as a function of incompatibility” // Journal of the Acoustical Society of Japan, Vol. 15, 1994, pp. 143-158.
84. Young S. et al. “The HTK Book (Version 3.5)”. Cambridge University Engineering Department, 2015.
85. Nefian A.V., Liang L.H., Pi X., Xiaoxiang X., Mao C., Murphy K., “A Coupled HMM for Audio-Visual Speech Recognition”, In Proc. International Conference ICASSP-2002, Orlando, USA, 2002, pp. 2013–2016.
86. Chu S., Huang T., “Multi-Modal sensory Fusion with Application to Audio-Visual Speech Recognition”, In Proc. Multi-modal Speech Recognition Workshop 2002, Greensboro, USA, 2002.
87. Rabiner L., Juang B., “Speech Recognition”, Chapter in Springer Handbook of Speech Processing, NY: Springer, 2008.
88. Potamianos G., Neti C., Luettin J., Matthews I., “Audio-Visual Automatic Speech Recognition: An Overview”, Chapter in “Issues in Visual and Audio-Visual Speech Processing”. MIT Press, 2005.
89. Bailly G., Perrier P., Vatikiotis-Bateson E., “Audiovisual Speech Processing”, Cambridge University Press, 2012, 506 p.
90. Stewart D., Seymour R., Pass A., Ming J., “Robust audio-visual speech recognition under noisy audio-video conditions” // IEEE Transactions on Cybernetics, Vol. 44, № 2, 2014, pp. 175-184.
91. Noda K., Yamaguchi Y., Nakadai K., Okuno H., Ogata T., “Audio-visual speech recognition using deep learning” // Applied Intelligence, Vol. 42, 2015, pp. 722-737.

92. Katsaggelos A.K., Bahaadini S., Molina R., “Audiovisual Fusion: Challenges and New Approaches” // Proceedings of the IEEE, Vol. 103, № 9, 2015, pp. 1635-1653.
93. Pierce, L. J. et al. // “Past experience shapes ongoing neural patterns for language”, Nat. Commun. 6:10073 doi: 10.1038/ncomms10073 (2015). http://www.nature.com/articles/ncomms10073
94. Wermke, K. et al. “Fundamental frequency variation within neonatal crying: Does ambient language matter?” // Speech, Language and Hearing, Vol. 19, Issue 4, 2016, pp. 211-217.
95. Дольник В. Непослушное дитя биосферы. «Петроглиф». СПб. 2011
96. Tomashenko N., Khokhlov Y., “Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing” // In Proc. INTERSPEECH 2014, Singapore, 2014, pp. 2997-3001.
97. Li, J., Yu, D., Huang, J.T., Gong, Y., “Improving wideband speech recognition using mixed-bandwidth training ” // Proceedings of the IEEE Spoken Language Technology Workshop (SLT), 2012, pp. 131–136.
98. Xue, J., Li, J., Yu, D., Seltzer,M., Gong, Y., “Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network”// Proc. ICASSP’2014, Florence, Italy, 2003. pp. 6409-6413.
99. Karafiat M., Burget L., Matˇejka P., Glembek O., ˇCernock´y J., “iVector-Based Discriminative Adaptation for Automatic Speech Recognition”// Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop, 2011, pp. 152-157.
100. Prudnikov A., Medennikov I., Mendelev V., Korenevskiy M., Khokhlov Yu., “Improving Acoustic Models For Russian Spontaneous Speech Recognition”// In Proc. 17th International Conference “Speech and Computer” SPECOM-2015, Athens, Greece, 2015, pp. 234-242.
101. Bengio Y., Ducharme R., Vincent P., “A neural probabilistic language model”// Journal of Machine Learning Research, Vjl. 3, 2003, pp. 1137-1155.
102. Schwenk H., Gauvain J.-L., “Training Neural Network Language Models On Very Large Corpora”// in Proc. Joint Conference HLT/EMNLP, October 2005, pp. 201-208.
103. Schwenk H., “Continuous space language models”// Computer Speech and Language Vol. 21, 2007, pp. 492–518.
104. Mikolov T., Karafia´t M., Burget L., C˘ernocky´ J., Khudanpur S., “Recurrent neural network based language model”// . INTERSPEECH 2010, Chiba, Japan, 2010, pp. 1045-1048.

УНИВЕРСИТЕТ ИТМО

Миссия университета – генерация передовых знаний, внедрение инновационных разработок и подготовка элитных кадров, способных действовать в условиях быстро меняющегося мира и обеспечивать опережающее развитие науки, технологий и других областей для содействия решению актуальных задач.

КАФЕДРА РЕЧЕВЫХ ИНФОРМАЦИОННЫХ СИСТЕМ

О кафедре

Кафедра речевых информационных систем (РИС) создана в 2011 году на факультете Информационных технологий и программирования (ФИТиП).

Организатором создания кафедры выступает «Центр речевых технологий» (www.speechpro.ru). Заведующий кафедрой – доктор технических наук Матвеев Юрий Николаевич.

Кафедра РИС обеспечивает подготовку докторантов, аспирантов и магистров. Для тех, кто имеет высшее образование, но хотел бы связать свое будущее с речевыми технологиями, имеются курсы дополнительного профессионального образования.

Обучение на кафедре

Кафедра «Речевые информационные системы» (базовая кафедра «Центра речевых технологий») Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики (ИТМО) в рамках направления 09.04.02 «Информационные системы и технологии» открывает прием в магистратуру по новой образовательной программе «Речевые информационные системы».

Срок обучения – 2 года. Обучение завершается защитой магистерской диссертации.

Целевая установка магистратуры – подготовка специалистов, способных участвовать в исследовательской и проектной работе в области речевых информационных технологий со специализацией в направлениях распознавания и синтеза речи, распознавания личностей по голосу, мультимодальной биометрии, в области проектирования и разработки информационных систем и программного обеспечения.

Область профессиональной деятельности выпускников кафедры РИС включает:

- исследование, разработка, внедрение речевых информационных технологий и систем;
- методы и алгоритмы цифровой обработки речевых сигналов;
- автоматизированные системы обработки речевых сигналов;

– программное обеспечение автоматизированных речевых информационных систем;

– системы автоматизированного проектирования программных и аппаратных средств для речевых информационных систем и информационной поддержки таких средств.

Объектами профессиональной деятельности выпускников кафедры РИС являются:

– информационные процессы, технологии, системы и сети, предназначенные для обработки, распознавания, синтеза речевых сигналов;

– инструментальное (математическое, информационное, техническое, лингвистическое, программное, эргономическое, организационное и правовое) обеспечение речевых информационных систем;

– способы и методы проектирования, отладки, производства и эксплуатации информационных технологий и систем в областях обработки, распознавания, синтеза речевых сигналов, телекоммуникации, связи, инфокоммуникации, медицины.

Широкий профиль подготовки, знание универсальных методов исследования и проектирования информационных систем, практические навыки работы с современным программным обеспечением – все это позволяет выпускникам кафедры найти работу в научных институтах и университетах, в фирмах, на производственных предприятиях, а также в коммерческих структурах. Учебный план предусматривает, в частности, следующие курсы:

– Информационные технологии:

– Системный анализ и моделирование информационных процессов и систем;

– Проектирование информационных систем;

– Организация проектирования и разработки программного обеспечения распределенных и встроенных систем;

– Тестирования программного обеспечения;

– Управление качеством разработки программного обеспечения.

Речевые технологии:

– Цифровая обработка сигналов;

– Цифровая обработка речевых сигналов;

– Математическое моделирование и теория принятия решений;

– Распознавание образов;

– Распознавание и синтез речи;

– Распознавание диктора (говорящего по голосу);

– Мультимодальные биометрические системы.

К преподаванию привлекаются ведущие специалисты «Центра речевых технологий», преподаватели Университета ИТМО, а также специалисты, работающие в известных научных организациях (СПИИРАН), а также производственных и коммерческих организациях.

Тампель Иван Борисович,
Карпов Алексей Анатольевич

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ РЕЧИ

Учебное пособие

В авторской редакции
Редакционно-издательский отдел Университета ИТМО
Зав. РИО Н.Ф. Гусарова
Подписано к печати 07.12.2017
Заказ № 4092
Тираж 100
Отпечатано на ризографе

Редакционно-издательский отдел
Университета ИТМО
197101, Санкт-Петербург, Кронверкский пр., 49