Е. Н. Мясникова

Объективное распознавание звуков речи

Москва
«Книга по Требованию»

УДК 62-63
ББК 30.6
Е11

Е. Н. Мясникова
Е11 Объективное распознавание звуков речи / Е. Н. Мясникова – М.: Книга по Требованию, 2012. – 150 с.

ISBN 978-5-458-29575-8

Библиотека по автоматике. Выпуск 242. Объективное распознавание звуков речи.

ISBN 978-5-458-29575-8

ВВЕДЕНИЕ

Задача объективного (автоматического) распознавания звуков речи, впервые поставленная и рассмотренная в 1943 г. [Л. 8], принадлежит к числу наиболее смелых и трудных научно-технических замыслов человека. Задача заключается в замене человека с его органом слуха устройством, которое, будучи способно воспринимать и распознавать звуки устной речи, осуществляло бы непосредственную связь человека с машиной. Это устройство должно выполнять соответствующие функции коры головного мозга человека, дающие возможность различать, «понимать» звуки речи. Примерами такого устройства служат диктофон-стенограф, автоматически печатающий звуки речи фонетическими знаками, и телефонный аппарат, автоматически набирающий номер с голоса.

Объективное распознавание звуков речи обычно относят к кругу задач распознавания образа, получивших за последние годы интенсивное развитие. Смежной задачей является осуществление автоматического чтения письменного текста машиной, т. е. распознавание зрительных образов [Л. 20]. За время, истекшее после появления уже упомянутой первой в мировой литературе работы [Л. 8], проблема привлекла к себе внимание многих исследователей ряда стран, причем за рубежом наиболее интенсивное ее развитие отмечается в США и в Японии.

Исследования, проведенные в этой области за последние годы, хотя и не ознаменовались еще полным успехом, но показали несомненную осуществимость задачи.

Проблему автоматического распознавания звуков речи можно рассматривать с различных позиций. Существуют два основных подхода, условно называемые психо-физическим и физико-техническим. При первом из них основное внимание уделяется исследованию процессов речеобразования и слухового восприятия, а при втором — исследованию акустических признаков звуков речи и разработке автоматически действующих электронных схем, способных выделять отличительные признаки речи. Мы исследуем этот вопрос, главным образом, с физико-технической точки зрения.

Первичными элементами звуков речи служат так называемые фонемы (гласные и согласные звуки речи). Фонемы входят в фонетический алфавит данного языка и обозначаются фонетическими знаками, которые в большинстве случаев совпадают с буквенными обозначениями. Эти звуки речи можно произносить по-разному в отношении громкости, длительности, тембра и т. д. Однако, несмотря на эти различия, человек способен определять принадлежность звуков к вполне определенным фонемам.

При произнесении, например, гласной фонемы *а* разными лицами, с разными голосами, протяжно или коротко, слушающий способен установить, что все это воспроизведение одной и той же фонемы *а*. Следуя распространенной в литературе терминологии [Л. 22], можно назвать фонетическую сущность звука, характеризующую принадлежность его к данному классу (т. е. к классу фонемы *а*), его образом. Распознавание фонемы есть установление класса, распознавание образа. Другие гласные (а также согласные) входят в другие классы, отличные по некоторым признакам, и дают другие образы.

За фонемами следуют более сложные фонетические образования: слоги, слова и фразы. Заметим, что сами фонемы могут быть раздроблены на более мелкие сегменты.

Главное значение с точки зрения распознавания речи имеют физические характеристики звуков речи, изуча-

емые акустикой. В последующих главах эти вопросы будут рассмотрены подробно, здесь же ограничимся несколькими вводными замечаниями.

Звук речи, распространяющийся в воздухе, есть совокупность звуковых волн, которую можно представить как наложение различных продольных упругих колебаний воздушной среды, причем под продольными колебаниями понимаются, как известно, периодические сжатия и расширения среды. Распределение упругих колебаний в пространстве представляет собой акустическое поле. Акустические волны распространяются со скоростью звука. В процессе распространения акустическое поле, вообще говоря, изменяется, происходит, например, ослабление колебаний, обусловленное расходимостью волн или их поглощением в среде.

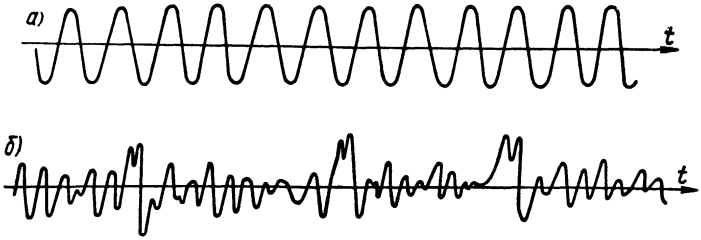

Рис. 1. Осциллограммы чистого тона и гласной: *а* — синусоидальное колебание; *б* — фонема *а*.

Если простой, чистый, протяжный тон можно изобразить в виде «синусоидальной» волны, т. е. волны, в которой колебательное смещение изменяется по синусоидальному закону (рис. 1, *а*), то звук речи дает волновую картину обычно весьма сложного вида. На рис. 1, *б* показана осциллограмма фонемы *а*, произнесенной протяжно. По координатным осям отложены смещение и время.

В качестве другого примера приведем оциллограмму согласной фонемы *с*. По своему составу звук *с* имеет шумовую природу: он аналогичен звуку выхода пара из сопла или собственному шуму лампового усилителя, который прослушивается в телефонных наушниках. Более подробное рассмотрение показывает, что звук *с* состоит из беспорядочно следующих друг за другом коротких импульсов разной интенсивности. Этот характер фонемы виден на осциллограмме, изображенной на рис. 2. Не

представляет труда отличить *а* от *с* по этим кривым. Нетрудно построить и электронную схему, которая автоматически отличала бы *а* от *с*. Если бы речь шла только об этих двух фонемах, то задачу объективного распознавания звуков речи можно было бы считать решенной. Однако фонем много, и некоторые из них имеют трудно уловимые отличия.

В акустике слово з в у к применяют часто в более широком смысле: звуком называют также электрические колебания, в которые с помощью микрофона преобразуются акустические колебания воздушной среды. Совершенно так же звуком называют ту модуляцию радиоволн в радиовещании и телевидении, которая осуществляет передачу речи и музыки. Звуком называют, наконец, и звукозапись (на магнитной ленте, кино-

Рис. 2. Осциллограмма согласной *с*.

пленке и др.). Во всех этих случаях, в которых мы имеем дело с преобразованием звуковых колебаний в электромагнитные, можно говорить о преобразовании фонемы из одного представления в другое. Если преобразование происходит без искажений, то вид осциллограмм не изменяется.

Основной задачей анализа звука служит определение акустических спектров. Сложную кривую периодического, длительного звука (такую, например, как осциллограмма фонемы *а*) можно разложить на синусоиды кратных периодов, т. е. осуществить разложение Фурье. Эти составляющие колебания отличаются друг от друга не только амплитудами, но и фазами; будучи вновь наложены одни на другие, они дадут первоначальную сложную фигуру. С п е к т р о м (или а к у с т и ч е с к и м с п е к т р о м) называют совокупность амплитуд составляющих синусоидальных колебаний с частотами, кратными основной, т. е. самой низкой. Получаются диаграммы, подобные спектру фонемы *а*, изображенному на рис. 3.

Компоненты спектра — спектральные линии — нельзя считать сколь угодно тонкими. В действительности каждая компонента имеет ширину, величина которой зависит от ряда причин и характеризует разрешающую способность анализа. В пределах ширины спектральной линии (рис. 4) дальнейший анализ затруднителен. Таким образом, частоту компоненты спектра можно ука-

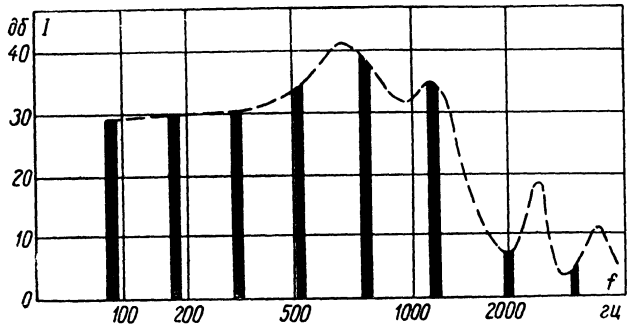

Рис. 3. Частотно-амплитудный спектр гласной *а*.

зывать только приближенно: обычно берут середину спектрального максимума.

Спектр фонемы *а* преимущественно линейчатый, т. е. состоит из отдельных спектральных линий. Характерно, что спектральные линии разделяются на группы. Из построения огибающей спектра (рис. 5) видно, что она имеет несколько максимумов. Области этих максимумов называются форма нтными, а частоты, соответствующие максимумам спектральных линий, — форма нтными часто тами. Звуки речи обладают форма нтами — т. е. спектральными областями, где амплитуды колебаний дают относительные максимумы.

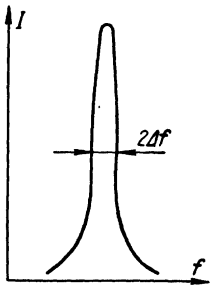

Рис. 4. Контур спектральной линии с полушириной Δf.

Происхождение формант обусловлено резонансами органа речи человека. Если изменить основной тон (пи тч) звука речи, что достигается управлением голосовыми связками, то изменяется частота не только наиболее низкой компоненты спектра, но и всех гармоник, часто́ты которых будут кратными.

Поэтому компоненты располагаются более или менее часто, смотря по тому, повышается или понижается тон. Однако огибающая, выделяющая формантные области, остается приблизительно той же, и формантные частоты сохраняются. В акустическом спектре другой фонемы (например, *о*) формантные области будут другими (см. § 1). Это важное обстоятельство указывает на то, что в формантах мы имеем некоторые отличительные признаки фонем.

Сравним гласную *а* с согласной *с*, акустический спектр которой приведен на рис. 6. Звук *с* — шумовой,

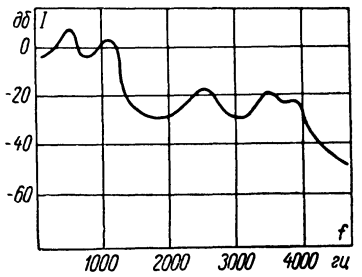

Рис. 5. Спектральная огибающая для гласной *а*.

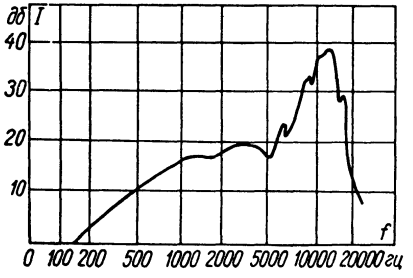

Рис. 6. Акустический спектр согласной *с*.

он представляет собой белый шум, в составе которого, как известно, присутствуют компоненты со всевозможными частотами, причем распределение амплитуд по частотам более или менее равномерно. Поэтому в акустическом спектре показываются не отдельные спектральные компоненты, сливающиеся вместе, а только огибающая, которая в данном случае имеет вид плавной линии.

Кроме частотных акустических спектров, можно использовать временны́е и амплитудные спектры, особенно удобные для характеристики случайных импульсных сигналов.

Временной спектр представляет собой зависимость вероятности временно́й длительности импульса от величины его временно́й длительности. Временно́й спектр не отражает амплитудных данных сигнала, и его обычно получают после клиппирования речевого сигнала. Под клиппированною речью понимают

речь с резкими ограничениями амплитуд, когда любое колебание и любой импульс превращается в стандартные прямоугольные импульсы постоянной высоты, сохраняющие различия длительности.

Пример образования клиппированного речевого сигнала показан на рис. 7. Временно́й спектр строят как диаграмму, выражающую распределение импульсов клиппированной речи по их длительностям (или по временам промежутков). Число импульсов данной длительности можно называть также частотой нулей, поскольку оно равно числу пересечений оси, на которой отклонение равно нулю. Таким образом, временно́й спектр выражает закон временного распределения частоты нулей клиппированной речи. Некоторые инва-

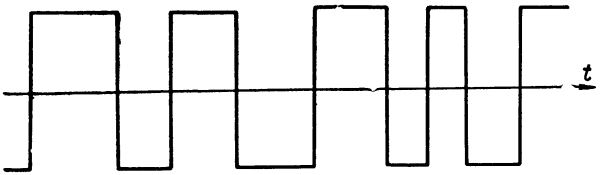

Рис. 7. Осциллограмма клиппированной речи.

риантные (лучше сказать квазиинвариантные) признаки клиппированной речи также используют для объективного распознавания звуков речи, которые в этом случае подлежат предварительному клиппированию.

Амплитудный спектр выражают диаграммой, дающей вероятность появления в сигнале компоненты с определенной амплитудой; эта вероятность наносится в зависимости от величины амплитуды. Амплитудные спектры особенно полезны для описания шумовых сигналов, характерных для многих согласных звуков речи. Например, на осциллограмме фонемы *с* наблюдается случайное чередование импульсов разной высоты, причем хаотическое распределение импульсов по осциллограмме дает равномерный сплошной частотный акустический спектр; амплитуды или выбросы подчиняются нормальному закону и амплитудный спектр изображается кривой Гаусса (рис. 8). При отклонениях от нормального статистического распределения амплитудные спектры изменяются.

Частотные, временны́е и амплитудные спектры — важнейшие объективные характеристики фонем. Однако

они не исчерпывают всех признаков, по которым можно распознавать фонемы и слова. Перечислим некоторые другие характеристики речевых сигналов.

Уровень звукового давления (или уровень интенсивности звука речи) не остается постоянным, а изменяется со временем. Об этом может свидетельствовать любая осциллограмма речи. Поэтому временно́е изменение уровня может давать некоторые признаки речевого сигнала. Чтобы получить уровни, необходимо сгладить отдельные колебания на малых участках осциллограммы и тем самым получить определенное отклонение, соответствующее амплитудному или среднеэффективному звуковому давлению или уровню интенсивности звука.

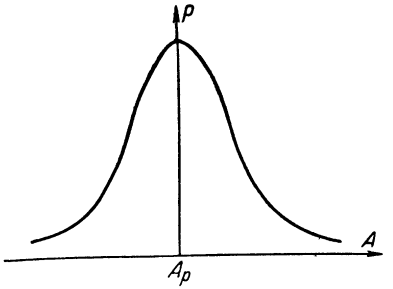

Рис. 8. Нормальное распределение амплитуд.
P — вероятность; A_p — наивероятнейшая амплитуда.

Изменение уровня звукового давления, принятое в качестве характеристики звука речи, в значительной степени зависит от громкости звука речи. Поэтому для использования такой характеристики необходима нормировка, т. е. приведение звука речи к какому-нибудь стандартному уровню громкости путем автоматической регулировки усиления. В некоторых случаях целесообразна компрессия, т. е. сужение динамического диапазона звука посредством нелинейного усиления. Изменение уровней во времени особенно полезно для распознавания не отдельных фонем, а целых слов.

Другой характеристикой служит изменение уровней звуковых давлений в зависимости от частоты и времени. Этот метод спектрального разложения звука, производимого в разные моменты времени, нашел себе широкое применение в анализе музыкальных фраз; применяется он и при анализе речи. Весьма существенны черты, характеризующие нарастание («атаку») и спадание звука.

Существуют и другие комбинированные способы осциллографирования и анализа звуков. Значительное

развитие получила, например, визуализация речи, производимая с помощью фиксации временных изменений частотного состава звука и интенсивности, причем изображение звуков получается на экране электронно-лучевой трубки. Поясним этот принцип на простой модели (дальнейшие подробности см. § 2).

Пусть частотный спектр изображается по вертикали в виде полоски, причем частота увеличивается вверх (рис. 9). Вторая полоска частотного спектра, соответствующая следующему промежутку времени, наносится

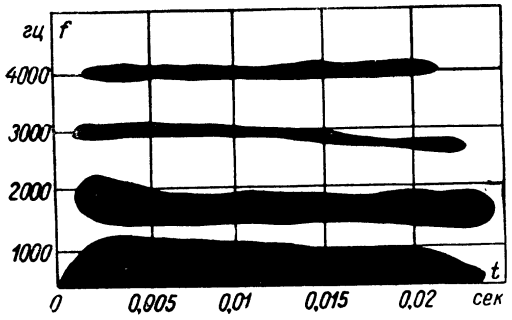

Рис. 9. Спектрограмма фонемы *а* при визуализации звуков речи.

рядом и так далее. Электронная трубка обладает послесвечением, так что изображения спектров остаются некоторое время на экране. За это время весь растр должен быть перекрыт. Что касается интенсивности, то она, изменяя яркость изображения, создает модуляцию по яркости. Развертка визуализированной речи должна быть достаточно быстрой для получения слитного изображения. Если для частотного анализа требуется известное время, и чем он подробнее, тем большее время занимает, то в данном методе частотный анализ производится с различной быстротой и подробностью. В ряде случаев для анализа достаточно использовать систему полосовых фильтров с интервалами, например, в треть октавы.

На рис. 10 изображены визуализированные фонемы *а* и *с*. По этим картинам можно легко отличить фонему *а*, обладающую вполне определенными спектральными максимумами и потому дающую горизонтальные

полосы, и фонему *с*, характеризуемую равномерным и широким заполнением поля. Особенно хорошо заметны смы́чки звуков речи — т. е. паузы, характерные для ряда согласных, например, *п* и *т* (см. § 2).

Визуализированную речь применяют для распознавания звуков речи глухими: зрение заменяет им отсутствующий слух.

С точки зрения бионики, объективное, автоматическое распознавание речи, осуществляемое средствами электроники, должно моделировать физиологические процессы распознавания речи человеком. Моделью уха служит микрофон плюс анализатор, выделяющий признаки фонем или слов, а моделью функциональной системы коры головного мозга, производящей распознавание,— электронная вычислительная машина, обладающая памятью и соответствующим образом программированная. В направлении реализации этой схемы и происходит современное развитие методов автоматического распознавания речи. Фонемы (и более сложные звуки речи) опознаются машинами как некоторые образы, которым соответствуют те или иные коды или знаки.

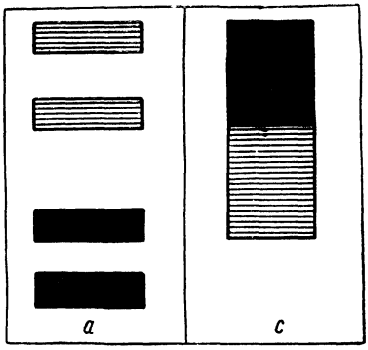

Рис. 10. Видеограммы фонем *а* и *с*.

Надо заметить, что фонетические знаки никак не отражают всего многообразия звучания даже отдельных фонем, зависящего от тональности, временно́й длительности и т. д. Кроме того, нельзя считать слог, а тем более слово, механическим соединением фонем, остающихся неизменными; наоборот, при образовании слов происходят существенные изменения составляющих фонем. В связи с этим возникает вопрос о том, закономерно ли говорить о фонемах как об определенных образах. Более правильно, быть может, называть фонемы псевдообразами или текущими образами. Нечего говорить о том, что такой характер фонем серьезно осложняет объективное распознавание речи.