Министерство образования и науки Российской Федерации
Ярославский государственный университет им. П. Г. Демидова
Кафедра инфокоммуникаций и радиофизики

А. И. Топников

Цифровая обработка речевых сигналов

Практикум

Ярославль
ЯрГУ
2018

УДК 004.939(076)
ББК 387-013я73
Т58

Рекомендовано
Редакционно-издательским советом университета
в качестве учебного издания. План 2018 года

Рецензент
кафедра инфокоммуникаций и радиофизики ЯрГУ

Топников, Артем Игоревич.

Т58 Цифровая обработка речевых сигналов : практикум / А. И. Топников ; Яросл. гос. ун-т им. П. Г. Демидова. – Ярославль : ЯрГУ, 2018. – 40 с.

Приводятся краткие теоретические сведения по ряду задач цифровой обработки речевых сигналов, а также даны практические задания, нацеленные на использование компьютерного моделирования.

Практикум предназначен для студентов, изучающих дисциплину «Цифровая обработка речевых сигналов». Материал также может быть использован при подготовке студентами курсовых и выпускных квалификационных работ.

УДК 004.939(076)
ББК 387-013я73

Введение

Обработка и передача речевых сигналов является важной составляющей современной радиотехники и ряда смежных областей. Доля данных, передаваемых в форме речевых сигналов, остается значительной, и их большая часть представлена в цифровом виде. Кроме того, речевые и звуковые сигналы являются важными компонентами видеосигналов. Велика роль и биометрических систем, использующих для идентификации личности речевые сигналы. Все это позволяет сделать вывод об актуальности исследований и знаний в области цифровой обработки речевых сигналов.

Методы и алгоритмы обработки речевых сигналов крайне многообразны, что вызвано кругом решаемых практических задач. Но существуют и общие термины, подходы и методы, получившие наиболее широкое применение. Этот базовый арсенал должен быть доступен всем специалистам в области радиотехники и цифровой обработки сигналов, поэтому именно на этих методах и подходах сделан акцент на страницах этого практикума.

Практикум посвящен вопросам детектирования голосовой активности, оценки качества и разборчивости речевых сигналов, шумоподавления и распознавания. Особое внимание уделено вопросам компьютерного моделирования. Внимательное ознакомление с краткими теоретическими сведениями, а также использование рекомендуемой литературы станет залогом успешного выполнения практических заданий, приведенных для каждой из рассматриваемых тем.

1. Детектирование голосовой активности

Задача детектирования голосовой (речевой) активности относится к числу наиболее часто встречающихся в области цифровой обработки речевых сигналов. Распознавание речевых сигналов и идентификация диктора начинаются с выделения фрагментов входного сигнала, содержащих речь. Многие методы шумоподавления и оценки качества также требуют сегментации входного сигнала на речевые и неречевые фрагменты. Особую роль играют детекторы голосовой активности (ДГА) в задаче сжатия речи – большинство современных кодеров содержат в своей структуре детектор голосовой активности.

Простейшие детекторы голосовой активности основываются на вычисление энергии отдельных фрагментов сигнала и сравнении этого значения с порогом. Логично предположить, что фрагменты, содержащие речь, обладают более высокой энергетикой, нежели фрагменты, содержащие паузы. Однако задача детектирования усложняется наличием в сигнале аддитивного шума. При значительном уровне шума простейшие детекторы речевой активности не в состоянии различить фрагменты, соответствующие некоторым низкоэнергетичным согласным звукам, и зашумленные паузы. Именно поэтому тематика, связанная с совершенствованием методов и алгоритмов детектирования голосовой активности, остается актуальной, а число научных публикаций по этой проблеме находится на высоком уровне.

Важным направлением повышения помехоустойчивости ДГА является поиск признаков речевых сигналов, позволяющих добиться наилучшей точности сегментации при низких отношениях сигнал/шум (ОСШ). Рассмотрим признаки, получившие наибольшее распространение.

Энергия. Для вычисления энергии фрагмента сигнала необходимо возвести в квадрат значения отдельных отсчетов и просуммировать получившиеся значения:

$$E = \sum_{i=1}^{N} E(i) = \sum_{i=1}^{N} s^2(i),$$

где $E(i)$ – энергия i-го отсчета $s(i)$ N – длина рассматриваемого фрагмента, который называют окном. Его длина выбирается исходя из особенностей решаемой задачи. В цифровой обработке речевых сигналов важным ориентиром является период стационарности сигнала. Наиболее часто используются окна длиной 10–30 мс.

Энергия Тигера. Появление этого признака связана с недостатками обычной энергии, которая не позволяет различать в сигнале участки, соответствующие низкоэнергетичным согласным звукам, и зашумленные паузы, обладающие близким уровнем энергии. Для решения этой проблемы предложена энергия Тигера (также известна как энергия Тигера-Кайзера, Teager-Kaiser energy), которая в дискретном случае для отсчета с номером i вычисляется следующим образом:

$$E(i) = s^2(i) - s(i-1) \cdot s(i+1).$$

Может возникнуть вопрос, связанный с вычислением энергии Тигера для первого и последнего отсчетов. Наиболее простое решение лежит в возведении этих отсчетов в квадрат без какого-либо вычитания, подразумеваемого приведенной формулой. В целом применение энергии Тигера вместо традиционной энергии и совместно с ней позволяет повысить точность детектирования голосовой активности при наличии в сигнале шумов значительной интенсивности.

Частота переходов через нуль. Этот признак также позволяет повысить точность работы ДГА, поскольку позволяет судить о частотных свойствах сигнала. Отвлекшись от речевых сигналов и рассмотрев простейшее гармоническое колебание, легко прийти к выводу, что, чем выше частота сигнала, тем большее число раз на фиксированном отрезке времени он пересечет нулевой уровень.

Реализация вычислений этого параметра зависит от выбранной среды программирования. Наиболее распространенный подход основан на сравнении полярности соседних отсчетов сигнала.

Кроме вычисления параметров (признаков) речевого сигнала, любой ДГА также содержит в себе некоторое устройство принятия решения, которое на основе их значений выносит решение,

содержит ли данный фрагмент сигнала речь или нет. Простейшие устройства принятия решения строятся на основе сравнения значений признаков с некоторым фиксированным или динамически меняющимся пороговым значением.

Таким образом, сочетание описанных выше признаков и порогового устройства позволяет построить несложный детектор голосовой активности, однако современные требования к системам обработки речевых сигналов заставляют искать более информативные признаки речевых сигналов, совершенствовать правила принятия решения, а также оптимизировать алгоритмы с целью снижения требуемых вычислительных затрат. В качестве примера современных ДГА могут служить стандартизованные методы детектирования, применяемые в области телекоммуникаций.

Например, работа детектора голосовой активности, применяемого в кодерах семейства G.729, стандартизованных Международным союзом электросвязи, основана на вычислении следующих характеристик входного сигнала:

– на разности уровня энергий,

– разности уровня энергий низкочастотного диапазона,

– искажении спектра,

– разности частоты переходов через ноль.

Работа этого алгоритма подразумевает постоянную оценку и учет параметров фонового шума, что позволяет значительно повысить устойчивость его работы в условии шумов.

Практические задания

1. Загрузите запись речевого сигнала. Разбейте сигнал на окна длиной 10–30 мс. Для каждого окна найдите значения таких параметров, как энергия, энергия Тигера, частота переходов через нуль. Сопоставьте на графике осциллограммы речевого сигнала и вычисленных параметров.

2. Постройте детектор речевой активности, основанный на простейших признаках речевого сигнала и сравнении их значений с пороговыми значениями. Обоснуйте выбор значений

порогов (для этого рекомендуется использовать гистограммы и диаграммы рассеяния параметров). Предложите модификации построенной схемы, обоснуйте их целесообразность.

3. Для построенного детектора голосовой активности и имеющейся речевой базы определите величину ошибок первого и второго рода для разных отношений сигнал/шум и разных типов шумов. Для зашумления используйте искусственно сгенерированный аддитивный белый гауссовый шум и записи реальных акустических шумов. Сделайте выводы.

4. Изучите модель детектора голосовой активности, применяемого в кодерах семейства G.729. Исследуйте точность работы детектора для разных отношений сигнал/шум и разных типов шумов. Для зашумления используйте искусственно сгенерированный аддитивный белый гауссовый шум и записи реальных акустических шумов. Проведите сравнение вашего детектора с данным.

2. Оценка качества

Одной из важнейших характеристик речевого сигнала в системах связи является качество. Стоит различать два близких понятия: качество и разборчивость. Под разборчивостью понимается степень, с которой человек способен правильно воспринимать отдельные структурные единицы речи. Качество же включает в себя значительное количество компонент и аспектов, что проявляется даже в том, как сложно дать конкретное и четкое определение этого понятия. Общепризнанные процедуры оценки качества искаженных речевых сигналов строятся на так называемых субъективных тестах, в рамках которых группе слушателей предлагается оценить качество звучащих речевых сигналов по некоторой, обычно пятибалльной, шкале. После проведения прослушивания полученные оценки усредняются. Итоговое значение называют качеством речи по шкале MOS (Mean Opinion Score). Сложность и затратность таких процедур, требующих непосредственного участия нескольких человек, привели к развитию объективных методов, заключающихся в математическом сравнении исходного (чистого) и искаженного сигналов.

Очевидно, что объективная оценка должна хорошо коррелировать с оценками субъективных прослушиваний. С этим связан особый интерес к изучению слуховой системы человека. Большинство современных объективных показателей качества речи учитывают психоакустические особенности человеческого слуха, благодаря чему достигается высокая достоверность оценки качества. В частности, при создании новых показателей качества принимается во внимании то, что:

1. Частотное разрешение уха неравномерно, т. е. частотный анализ акустических сигналов производится не в линейной шкале частот. Это можно смоделировать, используя банк полосовых фильтров, центральные частоты и полосы которых увеличиваются с ростом частоты. Полосы пропускания таких фильтров соответствуют так называемым критическим полосам.

2. Громкость связана с интенсивностью звуков нелинейным образом. Также учитывается, что воспринимаемая громкость изменяется в зависимости от частоты.

Рассмотрим наиболее распространенные методы, применяемые для оценки качества речевых сигналов. Все эти методы для своей работы требуют знания чистого (неискаженного) сигнала, который выступает в качестве своеобразного эталона, обладающего наилучшим качеством. Такие методы называются эталонными. В настоящее время активно развиваются и неэталонные методы, не требующие для оценки качества знания чистого сигнала, однако они обладают большей сложностью и, как правило, обеспечивают требуемый уровень достоверности лишь для ограниченного класса искажений или внешних условий.

Меры на основе отношения сигнал/шум. Общее отношение сигнал/шум, вычисленное для всего речевого сигнала, имеет низкую корреляцию с качеством, поэтому редко применяется в качестве объективного показателя. Большее распространение получило сегментное ОСШ (СегОСШ или segSNR, SNRseg), являющееся одной из простейших мер, применяемых для оценки работы алгоритмов кодирования или улучшения качества речи. Для правильного вычисления этой характеристики необходимо точное выравнивание сигналов на временной оси. Сегментное ОСШ во временной области определяется следующим образом:

$$segSNR = \frac{10}{M}\sum_{m=0}^{M-1}\lg\frac{\sum_{n=Nm}^{Nm+N-1}x^2(n)}{\sum_{n=Nm}^{Nm+N-1}\left(x(n)-y(n)\right)^2},$$

где $x(n)$ – исходный (чистый) сигнал, $y(n)$ – искаженный сигнал, N – длина окна (обычно составляет 15–20 мс), M – длина сигнала, выраженная в числе окон.

Основная потенциальная проблема, осложняющая вычисление СегОСШ, состоит в том, что его значение будет крайне низким для отрезков, содержащих паузы. Наличие пауз будет ухудшать общее среднее значение СегОСШ. Для решения данной проблемы можно не учитывать окна, содержащие паузы (для этого применяют детектор голосовой активности). Можно также не учитывать окна, для которых значение СегОСШ ниже

определенного порога. Обычно учитываемые значения СегОСШ ограничивают диапазоном [-10, 35] дБ, что позволяет избегать применения детектора речевой активности при вычислении среднего значения СегОСШ.

Сегментное ОСШ может быть также определено в частотной области. Такая характеристика называется частотно-взвешенным сегментным ОСШ и вычисляется следующим образом:

$$fwSNRseg = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} B_j \lg\left[\frac{X^2(m,j)}{\left(X(m,j)-Y(m,j)\right)^2}\right]}{\sum_{j=1}^{K} B_j},$$

где B_j – весовой коэффициент для j-й частотной полосы, K – число полос, M – длина сигнала, выраженная в числе окон, $X(m,j)$ – спектральная составляющая чистового сигнала в j-й частотной полосе в m-м окне, $Y(m,j)$ – спектральная составляющая искаженного сигнала в той же полосе и том же окне. Одним из главных преимуществ предложенного подхода по сравнению с вычислением сегментного ОСШ во временной области является возможность учета вклада частотных полос с разными весами. Число и ширина полос, как правило, выбираются с учетом особенностей восприятия речи слуховой системой человека. Зачастую вычисления осуществляются в критических полосах. Наборы весовых коэффициентов могут быть разными.

Меры качества на основе линейного предсказания. Линейное предсказание нашло широкое применение в задачах анализа и кодирования речевых сигналов. Используется оно и для оценки качества. Существует несколько методов, позволяющих оценить качество речевого сигнала на основе расстояния между наборами коэффициентов линейного предсказания, найденных для чистого и искаженного сигналов. В их основе лежит предположение, что на коротких временных участках речь может быть представлена моделирующим фильтром, состоящим из одних полюсов. Наиболее известными показателями качества, основанными на линейном предсказании, являются логарифмическое отношение правдоподобия (log-likelihood ratio, LLR) и мера Итакуро-Саито. Рассмотрим первый из методов.

Мера, основанная на логарифмическом отношении правдоподобия, определяется следующим образом:

$$LLR = d_{LLR}\left(a_x, \overline{a}_y\right) = \log\frac{\overline{a}_y^T R_x \overline{a}_y}{a_x^T R_x a_x},$$

где $\overline{a}_x$ – коэффициенты линейного предсказания для чистого сигнала, $\overline{a}_y$ – коэффициенты линейного предсказания для искаженного сигнала, R_x – автокорреляционная матрица чистого сигнала. Вычисление показателя осуществляется для отдельных окон, на которые разбивается сигнал. Затем значения, полученные для разных окон, за исключением 5 % наибольших, усредняются. Эта мера учитывает отклонения в расположении формантных максимумов сигналов. Чем ниже значение LLR, тем выше качество сигнала.

Взвешенный наклон спектра. Этот показатель качества основан на измерении взвешенной разности между наклонами спектра в критических полосах для двух сравниваемых сигналов. На первом шаге вычисления взвешенного наклона спектра (Weighted Spectral Slope, WSS) находятся энергии эталонного и оцениваемого сигналов в 25 критических полосах. Наклон спектра в каждой полосе для эталонного и оцениваемого сигналов вычисляется на основе следующих формул:

$$\Delta E_x(f) = E_x(f+1) - E_x(f)$$

$$\Delta E_y(f) = E_y(f+1) - E_y(f)$$

где $E_x(f)$ – энергия незашумленного сигнала-эталона $x(n)$ в критической полосе с номером f, $E_y(f)$ – энергия оцениваемого (зашумленного) сигнала $y(n)$ в критической полосе с номером f. Затем находятся пики $P(f)$ энергетических спектров в критических полосах, вычисленных на предыдущем этапе. Это позволяет вычислить веса:

$$W(f) = \frac{W_x(f) - W_y(f)}{2},$$

где

$$W_x(f) = \frac{20}{20 + E_{x,\max} - E_x(f)} \frac{1}{1 + P_x(f) - E_x(f)},$$

$$W_y(f) = \frac{20}{20 + E_{y,\max} - E_y(f)} \frac{1}{1 + P_y(f) - E_y(f)},$$

где $E_{x,\max}$ и $E_{y,\max}$ – максимальные значения $E_x(f)$ и $E_y(f)$ соответственно.

На основе вычисленных выше величин находится значение взвешенного наклона спектра WSS:

$$WSS = \frac{1}{N} \sum_{k=1}^{N} \left[\frac{\sum_{f=1}^{24} W(f) \left(\bar{\Delta} E_x(f) - \bar{\Delta} E_y(f) \right)^2}{\sum_{f=1}^{24} W(f)} \right],$$

где k – номер временного окна (все величины в скобках вычисляются отдельно для каждого окна), N – число временных окон, по которым осуществляется усреднение. Чем ниже значение WSS, тем лучше качество оцениваемого сигнала.

Показатель качества PESQ. Рассмотренные выше показатели пригодны для оценки качества речи, искаженной лишь ограниченным набором типов негативных воздействий, и не учитывают искажений, возникающих при передаче речевых сигналов по сетям связи. Искажения, вызванные потерей пакетов или работой кодеков, могут привести к неправильной оценке качества при использовании многих объективных мер. Для учета таких искажений в 1990-е годы были предложены специализированные критерии качества. В 2002 году исследовательской группой МСЭ-Т был проведен конкурс по выбору критерия, способного оценивать качество речевых сигналов, искаженных при передаче по сетям связи. В качестве замены прежней рекомендации P.861 была создана новая рекомендация P.862, основанная на мере качества PESQ (Perceptual Evaluation of Speech Quality), в то время как предыдущая рекомендация основывалась на мере качества PSQM (Perceptual Speech Quality Measure).

Кратко рассмотрим метод PESQ. На первом этапе осуществляется выравнивание уровня и полосовая фильтрация чистого (эталонного) и оцениваемого сигналов. Затем сигналы выравниваются по времени, чтобы скомпенсировать временные задержки, а потом поступают на банк фильтров, моделирующий работу слуховой

системы человека. Работа этих фильтров осуществляется в спектральной области. Получившиеся спектры эталонного и оцениваемого сигналов вычитаются друг из друга, таким образом вычисляется ошибка. В отличие от большинства методов оценки качества, которые не разделяют отрицательные и положительные ошибки, метод PESQ разделяет эти два вида ошибок, так как они по-разному влияют на ухудшения качества. Положительное значение ошибки свидетельствует о дополнительной аддитивной составляющей, вызванной шумом, а отрицательное – о том, что спектральная составляющая была полностью или частично подавлена. Исходя из разного восприятия двух типов искажений, положительные и отрицательные ошибки учитываются с разными весами. Значения, характеризующие искажения в локальной частотно-временной области, усредняются по частоте и времени. На основании информации о положительных и отрицательных ошибках рассчитываются два коэффициента: симметричный и асимметричный. На их основе вычисляется значение критерия качества PESQ:

$$PESQ = a_0 + a_1 \cdot d_{sym} + a_2 \cdot d_{asym},$$

где d_{sym} – коэффициент симметричных искажений, d_{asym} – коэффициент асимметричных искажений, $a_0 = 4{,}5$, $a_1 = -0{,}1$, $a_2 = -0{,}0309$. Максимальное значение показателя PESQ, соответствующее наивысшему качеству, равно 4,5. Стоит отметить, что данный показатель создавался прежде всего для оценки качества речи в IP-сетях и других телекоммуникационных системах, поэтому при его использовании в других приложениях корреляция его значений с субъективным восприятием может снижаться.

Комбинированный показатель качества. Так как в процессе подавления шума в речевом сигнале искажению подвергается не только сам сигнал, но и фон, то возникает необходимость в отдельной оценке качества сигнала и фона (остаточного шума). Именно эти две составляющие во многом определяют восприятие качества речевого сигнала в целом. Однако современные исследования свидетельствуют о том, что ни одна из двух основных составляющих качества, как и общее качество, не могут быть точно оценены при помощи каких-либо из существующих объ-

ективных показателей качества. Под точностью в данном случае подразумевается корреляция объективных метрик с субъективной оценкой группой слушателей.

В качестве основной тенденции, направленной на повышение точности методов объективной оценки качества, можно выделить создание комбинированных показателей качества. Основой для создания таких методов служат результаты субъективной оценки, существующие объективные критерии качества, а также теория регрессионного анализа. С помощью этой теории, а также набора экспериментальных данных качество сигнала, фона или общее качество могут быть выражены как линейная комбинация значений нескольких показателей качества. Рассмотрим одну из реализаций данного подхода – трехкомпонентный комбинированный показатель качества. Согласно этому методу качество оценивается тремя коэффициентами: Csig (качество речевого сигнала), Cbak (качество фона), Covl (общее качество). Каждый из этих коэффициентов вычисляется как линейная комбинация значений определенных объективных показателей качества. Для нахождения трех коэффициентов необходимо первоначально вычислить значения восьми объективных показателей качества, а затем подставить их в следующие формулы:

$$Csig = 3,093 - 1,029 \cdot LLR + 0,603 \cdot PESQ - 0,009 \cdot WSS;$$

$$Cbak = 1,634 + 0,478 \cdot PESQ - 0,007 \cdot WSS + 0,063 \cdot segSNR;$$

$$Covl = 1,594 + 0,805 \cdot PESQ - 0,007 \cdot WSS - 0,512 \cdot LLR.$$

Значения данных коэффициентов для большинства речевых сигналов изменяются в пределах от 0 до 5, однако в некоторых случаях их значения могут выходить за пределы указанного диапазона. Значения приведенных коэффициентов демонстрируют высокую корреляцию с результатами субъективных тестов для широкого набора шумовых условий.

Практические задания

1. Используя готовые реализации объективных методов, произведите численную оценку качества зашумленных речевых сигналов. В качестве источника зашумления используйте искусственно сгенерированный аддитивный белый гауссовский шум, а также записи реальных акустических шумов из специализированной фонотеки.

2. Оцените корреляцию значений показателей качества в широком диапазоне отношений сигнал/шум как для аддитивного белого гауссовского шума, так и для одного из видов реальных шумов. Для каждой аудиозаписи речевого фрагмента и фиксированного типа шума необходимо сгенерировать не менее пятидесяти реализаций шума. Результаты представьте в форме таблиц (табл. 1) и соответствующих диаграмм рассеяния (рис. 1). Сделайте выводы о степени статистической взаимосвязи показателей качества речевых сигналов.

Таблица 1

Показатели	*Показатель 1*	*Показатель 2*	. . .	*Показатель N*
Показатель 1	1			
Показатель 2		1		
. . .			1	
Показатель N				1

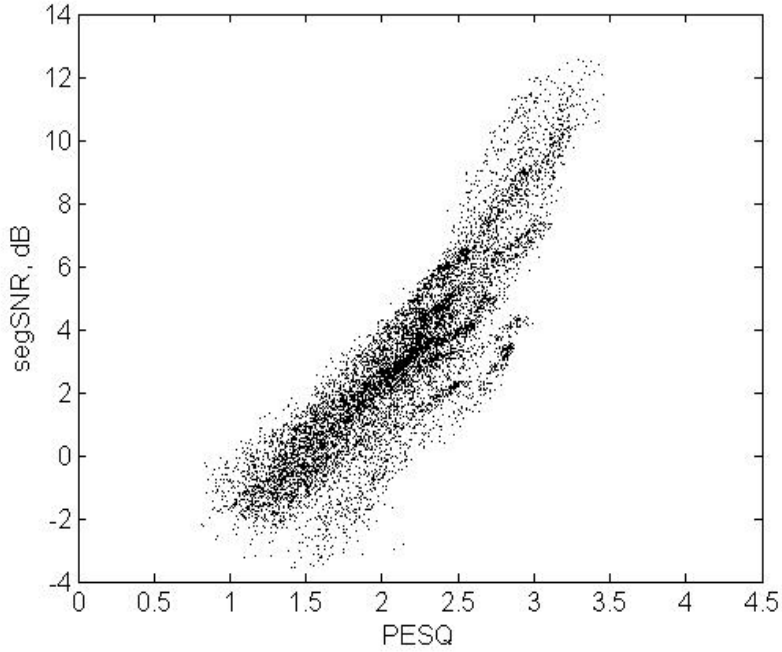

Рис. 1. Диаграмма рассеяния
для пары показателей segSNR и PESQ

3. Исследуйте возможность использования математического аппарата линейной регрессии для оценки значений одного показателя качества по известным значениям другого. Для численной характеристики точности предлагается использовать значения средней абсолютной ошибки, корня из среднеквадратической ошибки и относительной ошибки, выраженной в процентах. На основе полученных результатов сделайте соответствующие выводы.

3. Оценка разборчивости

Наряду с качеством, не менее важной характеристикой речевых сигналов в радиотехнических системах связи является разборчивость. Речевой сигнал в первую очередь является носителем смысловой информации. Способность человека воспринимать отдельные единицы речи характеризуется разборчивостью. Наличие шумов и других типов искажений в радиотехнических системах способно значительно снизить разборчивость речи и тем самым свести способность к ее восприятию к минимуму. Именно поэтому оценка разборчивости речи является важной задачей в области радиотехники и цифровой обработки речевых сигналов. Оценка разборчивости важна как на стадии разработки методов и подсистем обработки речи в радиотехнических системах, так и при их эксплуатации.

Методы оценки разборчивости речевых сигналов, так же как и методы оценки качества речи, можно разделить на субъективные и объективные. В первой группе методов разборчивость речи характеризуется непосредственной способностью слушателей воспринимать речевые фрагменты. Для данной группы методов характерна высокая точность, однако они требуют непосредственного привлечения дикторов и слушателей, что приводит к существенным временным и финансовым затратам. Объективные критерии разборчивости позволяют производить оперативную оценку разборчивости с минимальными затратами, однако в большинстве случаев учитывают влияние на разборчивость лишь ограниченного числа негативных факторов.

При проведении субъективной оценки разборчивости дикторы произносят в умеренном темпе слова или фразы, а слушатели (эксперты) записывают на бумагу или выделяют подчеркиванием услышанные речевые фрагменты. Отношение правильно услышанных (записанных) речевых единиц к общему числу произнесенных, выраженное в процентах, является численной характеристикой разборчивости. Для оценки разборчивости могут использоваться звуки, слоги, слова или фразы. В зависимости от типа используемого речевого материала выделяют звуковую, слоговую, словесную и фразовую разборчивость.

Для одного и того же набора условий разборчивость перечисленных типов, выраженная в процентах, будет иметь разные значения. При прочих равных условиях численное значение фразовой разборчивости будет всегда больше словесной, словесная – слоговой и т. д. Это связано с разной степенью неопределенности и способностью человека угадывать элементы речи благодаря наличию контекста и определенных закономерностей речи. Выделенные типы разборчивости взаимосвязаны друг с другом и при необходимости возможен пересчет. Выбор материала для проведения исследования в первую очередь определяется удобством проведения экспериментов.

Различные методы и стандарты регламентируют особенности проведения экспериментов по субъективной оценке разборчивости. Диктор должен обладать хорошей дикцией. Желательно, чтобы его голос был незнаком экспертам. Слушатели должны обладать здоровой слуховой системой (поэтому зачастую налагаются ограничения на возраст); наличие требований к специальной подготовке зависит от метода. Число слушателей также определяется используемым методом или стандартом.

Особое внимание методов оценки разборчивости уделено материалу для зачитывания – артикуляционным таблицам. Выбор речевого материала должен учитывать особенности языка, максимально полно представлять его особенности, охватывать всевозможные сочетания речевых единиц. Например, при определении словесной разборчивости в артикуляционную таблицу должны быть включены слова с разным числом слогов и разным расположением ударений (на первый слог, на второй и т. д.). Совокупность зачитываемых слов должна максимально полно содержать набор слогов, встречающихся в конкретном языке. В ряде методов экспертам предлагается выбрать одно из предложенных слов.

В целом для субъективных методов оценки разборчивости характерна высокая степень достоверности при высоком уровне затратности, вызванном необходимостью участия относительно большого числа людей и сложностью соблюдения других требований.

Для решения большинства практических задач используются объективные методы оценки разборчивости, достоверность которых оценивается путем сравнения с результатами субъективных тестов. Если результаты объективной оценки разборчивости для широкого круга возможных условий показывают высокую корреляцию с оценками, полученными одним из субъективных методов, то данная объективная мера разборчивости может считаться достоверной и использоваться для оценки разборчивости в процессе разработки или эксплуатации систем обработки и передачи речевых сигналов.

За последние десятилетия создано большое количество объективных методов оценки разборчивости. Каждый метод создается для решения определенных задач и имеет свои ограничения, так как учитывает влияние на разборчивость речевых сигналов лишь ограниченного круга негативных факторов. Все объективные показатели разборчивости можно разделить на формантные (AI, SII, fAI, SNR loss), модуляционные (STI, RASTI, STIPA, STITEL), эмпирические (%ALcons, C50) и прочие. Рассмотрим методы, относящиеся к первым двум группам, наиболее востребованным в области радиотехники.

Метод AI. Наиболее известным и широко распространенным формантным методом оценки разборчивости является метод AI (Articulation Index или индекс артикуляции). Теоретические основы, на которых базируется данный метод, были заложены в первой половине XX века исследованиями Х. Флетчера, Д. Колларда, Н. Френча и Д. Стеинберга. В отечественной научной литературе развитие этого подхода связано с именами Н. Б. Покровского, Ю. С. Быкова и М. А. Сапожкова. Идея метода состоит в разбиении спектра сигнала на некоторое количество полос и измерении ОСШ в каждой из них. Данный метод основывается на предположении, что каждая полоса вносит свой независимый вклад в общую разборчивость речевого сигнала. Параметры полос выбираются так, чтобы каждая из них имела равный вклад в общую разборчивость.

Рассмотрим вычисление показателя AI в упрощенном виде. На первом этапе метода находятся значения ОСШ в каждой

из 20 полос (число полос может быть и иным). Затем полученное значение ОСШ в определенной полосе масштабируется таким образом, чтобы уровню более 30 дБ соответствовало значение 0,05. Получившееся значение (индивидуальный индекс разборчивости) характеризует разборчивость в конкретной полосе. Индекс артикуляции определяется как сумма индивидуальных индексов разборчивости и изменяется в пределах от 0 до 1. Значения AI ниже 0,3 соответствуют плохой разборчивости; от 0,3 до 0,5 – удовлетворительной; от 0,5 до 0,7 – хорошей; выше 0,7 – очень хорошей. Значения AI имеют нелинейную связь со словесной разборчивостью, выраженной в процентах. Однако в большинстве случаев на практике для характеристики уровня разборчивости речевых сигналов достаточно указания значения AI.

Метод SII. Дальнейшее развитие идей, лежащих в основе создания метода AI, привело к созданию метода SII (Speech Intelligibility Index или индекс разборчивости речи). Этот метод также подразумевает измерение ОСШ в отдельных полосах с последующим суммированием значений, характеризующих разборчивость в отдельных полосах. Главное же отличие от критерия AI состоит в том, что вместо набора из 20 полос в данном подходе используются 4 различных набора полос:

– критические полосы;

– третьоктавные полосы;

– критические полосы, обладающие равным вкладом в разборчивость;

– октавные полосы.

Значение критерия SII также изменяется от 0 до 1.

Описанные выше методы могут применяться для оценки разборчивости речи в широком круге задач, но степень их достоверности может существенно варьироваться. Большинство современных методов шумоподавления являются нелинейными, поэтому зашумленный речевой сигнал в процессе обработки подвергается нелинейным искажениям. Без учета особенностей искажений, вносимых методами шумоподавления, невозможно создать объективную меру разборчивости речи, обладающую

высокой достоверностью. Примерами методов, позволяющих достаточно точно оценивать разборчивость речевых сигналов на выходе методов (систем) шумоподавления, являются методы SNR loss и fAI.

Метод SNR loss. Рассмотрим показатель разборчивости SNR loss. Пусть речевой сигнал зашумлен аддитивным шумом:

$y(n) = x(n) + d(n)$,

где $x(n)$ – чистый речевой сигнал, а $d(n)$ – шум. После обработки алгоритмом шумоподавления зашумленный сигнал $y(n)$ преобразуется в сигнал $z(n)$, являющийся оценкой чистого сигнала $x(n)$.

С помощью оконного преобразования Фурье вычисляется частотно-временной образ зашумленного речевого сигнала:

$$Y(k,m) = \sum_{n=0}^{N-1} y(mR+n) \cdot h(n) e^{-\omega_k n},$$

где k = 0,1,…, N – 1 – номер частотного интервала (бина), N – длина окна, $\omega_k = 2\pi k/N$ – частота, соответствующая k-му бину, m – номер временного интервала (окна), R – коэффициент, определяющий перекрытие соседних окон, $h(n)$ – оконная функция Хэмминга, n – временной индекс внутри окна. Аналогичное преобразование выполняется для незашумленного (эталонного) сигнала.

Путем перемножения амплитудного спектра $|Y(k,m)|$ и 25 гауссовских окон и возведения в квадрат находится энергия сигнала в критических полосах. В результате этой операции получаем частотно-временной образ сигнала:

$$Y(j,m) = X(j,m) + D(j,m),$$

где j – номер критической полосы, $X(j,m)$ – спектральная составляющая чистого сигнала для j-й полосы и m-го временного интервала, $D(j,m)$ – аналогичная составляющая для шумового сигнала.

Для каждой j-й полосы каждого m-го кадра вычисляется следующая величина:

$$L(j,m) = SNR_X(j,m) - SNR_Z(j,m) =$$

$$= 10 \cdot \lg \frac{X^2(j,m)}{D^2(j,m)} - 10 \cdot \lg \frac{Z^2(j,m)}{D^2(j,m)} = 10 \cdot \lg \frac{X^2(j,m)}{Z^2(j,m)},$$

где $SNR_X(j,m)$ – входное отношение сигнал/шум (ОСШ) в полосе j и интервале m, $SNR_Z(j,m)$ – ОСШ после обработки в полосе j и интервале m, $Z(j,m)$ – j-я составляющая спектра обработанного сигнала, вычисленного с учетом критических полос для m-го временного интервала. Очевидно, что при $Z(j,m) = X(j,m)$ величина $L(j,m) = 0$. В целом значение $L(j,m)$ может быть как положительным, так и отрицательным.

Значения $L(j,m)$ ограничиваются в пределах определенного диапазона $[-SNR_{\text{lim}}, SNR_{\text{lim}}]$:

$$L_{\text{lim}}(j,m) = \min\left(\max\left(L(j,m), -SNR_{\text{lim}}\right), SNR_{\text{lim}}\right).$$

Полученные на предыдущем этапе значения $L_{\text{lim}}(j,m)$ масштабируются на интервал [0, 1]:

$$SNR_{LOSS}(j,m) = \begin{cases} -\dfrac{C_-}{SNR_{\text{lim}}} L_{\text{lim}}(j,m), & \text{если } L_{\text{lim}}(j,m) < 0 \\ \dfrac{C_+}{SNR_{\text{lim}}} L_{\text{lim}}(j,m), & \text{если } L_{\text{lim}}(j,m) \geq 0, \end{cases}$$

где C_- и C_+ – параметры масштабирующей функции.

Для каждого окна в отдельности осуществляется усреднение $SNR_{LOSS}(j,m)$ по всем критическим полосам:

$$fSNR_{LOSS}(m) = \frac{\sum_{j=1}^{K} W(j) \cdot SNR_{LOSS}(j,m)}{\sum_{j=1}^{K} W(j)},$$

где $W(j)$ – весовые коэффициенты, учитывающие психоакустические закономерности восприятия речевых сигналов.

Вычисляется среднее значение $\overline{SNR}_{LOSS}$ путем усреднения $fSNR_{LOSS}(m)$ по всем окнам:

$$\overline{SNR}_{LOSS} = \frac{1}{M} \sum_{m=0}^{M-1} fSNR_{LOSS}(m),$$

где M – число окон, на которые разделен сигнал. Получившаяся величина характеризует разборчивость сигнала. Она изменяется в интервале от 0 до 1. Нулевое значение соответствует идеальной разборчивости, а единичное – полному отсутствию разборчивости.

В отличие от многих других объективных критериев разборчивости речи, критерий SNR loss изначально создавался для оценки разборчивости зашумленной речи, обработанной системой шумоподавления. Учет нелинейного характера обработки существующих методов шумоподавления позволяет достичь высокой достоверности оценки разборчивости речевых сигналов, обработанных при помощи современных методов шумоподавления.

Метод fAI. Другим современным методом оценки разборчивости речи на выходе систем шумоподавления является fAI, базирующийся на широко распространенном объективном показателе разборчивости AI, но, в отличие от него, учитывающем нелинейный характер преобразований, производимых алгоритмами шумоподавления. Данный метод адаптирован для оценки разборчивости речи на выходе методов шумоподавления, основанных на применении к амплитудному спектру сигнала функции коррекции спектра, значения которой изменяются в пределах от 0 до 1.

Рассмотрим вычисление показателя разборчивости fAI. Опустим вычисление частотно-временных образов сигналов, так как оно производится аналогично методу SNR loss и перейдем к рассмотрению последующих вычислений. Для каждой *j*-й полосы каждого m-го окна вычисляются следующие величины:

$$SNR_X(j,m) = 10 \cdot \lg \frac{X^2(j,m)}{D^2(j,m)},$$

$$\overline{SNR_X}(j,m) = 10 \cdot \lg \frac{Z^2(j,m)}{D^2(j,m)},$$

где $SNR_X(j,m)$ – входное ОСШ в полосе j и окне m, $\overline{SNR_X}(j,m)$ – ОСШ после обработки в полосе j и окне m, $Z(j,m)$ – j-я составляющая спектра обработанного сигнала, вычисленного с учетом критических полос для m-го временного окна, $X(j,m)$ – j-я составляющая спектра входного сигнала, вычисленного с учетом критических полос для m-го временного окна.

Затем вычисляется следующая величина:

$$fSNR(j,m)=\begin{cases}\dfrac{\min(SNR_X(j,m),\overline{SNR_X}(j,m))}{SNR_X(j,m)}, & \text{если } SNR_X(j,m)\geq SNR_L\\ 0 \quad \text{в остальных случаях}\end{cases},$$

где SNR_L – задаваемое минимально разрешенное ОСШ. Использование данной величины гарантирует, что будут использоваться только те компоненты, для которых составляющие спектра обработанного сигнала больше шумовых составляющих. Исследования показали, что наибольшая корреляция значений критерия fAI с результатами субъективных тестов достигается при SNR_L = 11 дБ.

Значение показателя fAI для *m*-го окна вычисляется следующим образом:

$$fAI=\frac{1}{\sum_{j=1}^{M}W_j}\cdot\sum_{j=1}^{M}W_j\cdot fSNR_j,$$

где M – число критических полос. После этого значения fAI, полученные для отдельных временных окон, усредняются. Получившееся на заключительном этапе метода значение может изменяться в пределах от 0 до 1. Чем выше значение fAI, тем выше разборчивость сигнала на выходе метода шумоподавления. Связь значения критерия fAI с разборчивостью *I*, выраженной в процентах, определяется следующим выражением:

$$I=\left(1-10^{-fAI\cdot P/Q}\right)^2,$$

где P = 27,5; Q = 8,4. Исходя из этого выражения, можно сделать вывод, что высокий уровень разборчивости (>90 %) соответствует значениям fAI, превышающим 0,5.

Метод STI. Кроме формантных методов широкое применение получил модуляционный метод STI (Speech Transmission Index или индекс передачи речи) и его модификации. Главное его отличие состоит в том, что для измерения разборчивости используется искусственно синтезированный сигнал. В качестве модели речевого сигнала выступает шумоподобный сигнал, имеющий

равномерное распределение спектральной плотности в каждой из семи октавных полос, покрывающих суммарно частотный интервал от 125 до 8000 Гц. Шумоподобный сигнал модулируется гармоническим сигналом низкой частоты (от 0,63 до 12,5 Гц). Для этого используется амплитудная модуляция с коэффициентом модуляции, близким к единице. При воздействии на сигнал аддитивного шума и ряда других негативных факторов сигнал искажается, при этом уменьшается значение коэффициента модуляции. Таким образом, коэффициент модуляции имеет взаимосвязь со степенью разборчивости. Значение критерия STI находится как взвешенное среднее коэффициентов модуляции, оцененных для разных октавных полос и частот модулирующего колебания. Значение критерия изменяется от 0 до 1. Отличная степень разборчивости соответствуют значениям STI>0,75. Для сопоставления результатов, полученных с использованием различных объективных методов оценки разборчивости, на основе критерия STI введена общая шкала разборчивости (Common Intelligibility Scale, CIS). Взаимосвязь CIS и STI описывается следующим выражением:

$$CIS = 1 + \log(STI).$$

Также объективный критерий STI послужил основой для создания множества других методов, предназначенных для оценки разборчивости в разных задачах (например, методы RASTI, STIPA, STITEL).

Практические задания

1. Используя речевую базу, проведите сравнительный анализ формантных показателей разборчивости. Для искажения речевых сигналов используйте искусственно сгенерированный аддитивный белый гауссовый шум, а также записи реальных акустических шумов.

2. Используя доступную литературу и другие источники, изучите отечественные версии формантного метода. Проведите их сравнительный анализ.

3. Изучите работу модуляционных методов оценки разборчивости, используя для моделирования искажений речевых сигналов импульсные характеристики, измеренные для различных помещений. Прослушайте получившиеся сигналы. Проанализировав численные значения показателя STI, сделайте выводы.

4. Используя одну из существующих методик, проведите субъективную оценку разборчивости речевых сигналов. Сопоставьте полученные результаты с результатами объективной оценки. Сделайте выводы.

5. Сформулируйте основные недостатки и ограничения изученных методов, предложите пути их совершенствования.

6. Возможна ли замена искусственно синтезированного сигнала в методе STI и его модификациях на реальный речевой сигнал? Если возможна, предложите необходимые изменения в структуре метода. Назовите достоинства и недостатки такой модификации.

4. Подавление шума

Решение большинства практических задач в области цифровой обработки сигналов осложняется присутствием шумов. И чем выше уровень шума и, соответственно, ниже отношение сигнал/шум, тем ниже качество и разборчивость сигналов, точность систем распознавания речевых сигналов и идентификации дикторов. Именно поэтому шумоподавление – один из наиболее распространенных видов предобработки в системах анализа и обработки речевых сигналов.

Наибольшее распространение получили методы, осуществляющие подавление шума в спектральной области. Значительный прогресс в их развитии наблюдался в 80–90-е годы XX века, однако до сих пор их можно считать базовыми, обязательными для изучения. Именно с них нужно начинать знакомство с задачей шумоподавления и только потом переходить к изучению методов шумоподавления, основанных на применении методов машинного обучения, ставших в последние годы наиболее перспективными.

Сформулируем задачу шумоподавления. Пусть речевой сигнал $x(t)$ искажается аддитивным шумом $n(t)$ тогда зашумленный сигнал $y(t)$ может быть записан следующим образом:

$$y(t) = x(t) + n(t).$$

Аналогичное выражение можно записать и для спектральной области:

$$Y_k = X_k + N_k,$$

где Y_k, X_k, N_k – соответствующие спектральные образы сигналов $y(t)$, $x(t)$, $n(t)$, а k – номер частотного интервала.

В общем виде задача шумоподавления может быть сформулирована как получение оптимальной по некоторому критерию оценки сигнала $x(t)$ (или его спектрального образа X_k) по наблюдаемому зашумленному сигналу $y(t)$ (или его спектральному образа Y_k). При этом, как правило, делаются некоторые предположения о статистических свойствах шума $n(t)$. Наиболее часто предполагается, что шум имеет нормальное распределение с нулевым математическим ожиданием.

Стоит отметить, что предположения о стационарности речевого сигнала и шума справедливы лишь для небольших интервалов времени (обычно, 10–30 мс.), поэтому при реализации методов шумоподавления используют разбиение сигнала на окна. Для упрощения обозначений не будем указывать в формулах индекс, отвечающий за номер окна.

Оценку спектра сигнала можно осуществлять отдельно для амплитудной и фазовой составляющих. В большинстве методов подавления аддитивного шума в речевых сигналах в качестве оценки фазового спектра незашумленного сигнала используется фазовый спектр зашумленного сигнала. Поэтому задачу шумоподавления можно свести к поиску оценки амплитудного спектра незашумленного сигнала Z_k. В спектральной области она может быть получена путем поэлементного перемножения амплитудного спектра зашумленного сигнала R_k и так называемой функции коррекции спектра (ФКС) G_k:

$$Z_k = G_k \cdot R_k .$$

Стоит отметить, что термин «функция коррекции спектра» не является общепринятым и ряд исследователей предпочитают использовать наименования, более близкие к англоязычному термину «гейн-функция» (от англ. gain-function), или «функция усиления».

Функция коррекции спектра обычно является функцией от априорного SNR_k^{prio} или апостериорного SNR_k^{post} отношения сигнал/шум, которые можно задать следующим образом:

$$SNR^{prio} \quad \frac{E\{A\}}{E\{D\}}$$

$$SNR_k^{post} = \frac{R_k^2}{E\{D_k^2\}},$$

где $E\{A_k^2\}$ – значение спектральной плотности чистого сигнала в полосе с номером k, $E\{D_k^2\}$ – значение спектральной плотности шума в полосе с номером k.

Спектральные свойства шума могут быть оценены на участках сигнала, не содержащих речь. Это могут быть паузы в начале сигналов или неречевые фрагменты, выделенные детектором

голосовой активности. Для оценки априорного отношения сигнал/шум часто используют метод прямого принятия решения, позволяющий оценить априорное отношение сигнал/шум на основе известного апостериорного.

К числу наиболее широко распространенных функций коррекции спектра можно отнести следующие:

ФКС метода спектрального вычитания:

$$G_{CB}\left(SNR_k^{post}\right)=\begin{cases}\sqrt{1-\dfrac{\beta}{SNR_k^{post}}} & при\ 1-\dfrac{\beta}{SNR_k^{post}}>0,\\ 0 & иначе,\end{cases}$$

где β – параметр алгоритма.

ФКС Винера:

$$G_{Винера}\left(SNR_k^{prio}\right)=\frac{SNR_k^{prio}}{1+SNR_k^{prio}}.$$

ФКС минимальной среднеквадратической ошибки кратковременного амплитудного спектра (MMSE-STSA, minimum mean square error short-time spectral amplitude):

$$G_{MMSE}\left(SNR_k^{prio},SNR_k^{post}\right)=\frac{\sqrt{\pi}}{2}\frac{\sqrt{\nu_k}}{SNR_k^{post}}\cdot {}_1F_1\left(-0,5;1;-\nu_k\right),$$

где $\nu_k=\dfrac{SNR_k^{prio}}{1+SNR_k^{prio}}\cdot SNR_k^{post}$, ${}_1F_1(a;b;x)$ – вырожденная гипергеометрическая функция.

Известна также группа методов, в которых функция коррекции спектра имеет вид бинарной маски:

$$G_{bin}=\begin{cases}1, & H_1\\ 0, & H_0,\end{cases}$$

где H_1 – гипотеза, состоящая в том, что в некоторой частотно-временной точке спектрального представления сигнала априорное отношение сигнал/шум превышает некоторый порог (обычно выбирается значение 0 дБ), а H_0 – гипотеза, состоящая в том, что априорное отношение сигнал/шум ниже порога. Таким образом зануляются спектральные составляющие, в которых уровень шума превышает уровень полезного сигнала.

Практические задания

1. Выберите один из методов шумоподавления и изучите его работу. Сопоставьте осциллограммы и спектрограммы входного и выходного сигналов при разных уровнях отношения сигнал/шум на входе.

2. С использованием объективных показателей качества и разборчивости проведите сравнительный анализ предложенных методов шумоподавления. Для зашумления сигналов используйте аддитивный белый гауссовский шум. Воспроизводя и прослушивая сигналы, сравните артефакты, вносимые методами шумоподавления. Повторите исследование, производя зашумление сигналов записями реальных акустических шумов.

3. Реализуйте модификацию одного из предложенных методов шумоподавления, осуществляя оценку спектральных характеристик шума с использованием детектора голосовой активности. Проведите сравнение с исходным методом.

4. Реализуйте метод шумоподавления на основе бинарных масок. Используя объективные показатели качества и разборчивости, подберите величину порога. Обоснуйте выбор показателей качества и разборчивости. Прослушайте речевые сигналы до и после обработки, опишите артефакты, вносимые методом шумоподавления.

5. Распознавание речевых сигналов

Современный уровень систем распознавания речи (преобразования речевого сигнала в текст) является результатом развития методов машинного обучения. Однако эти методы применяются совместно с подходами и методами, разработанными во второй половине XX века. В основе многих систем распознавания изолированных слов лежит схема, основанная на сравнении распознаваемого образа с эталоном (рис. 2). Сигнал от микрофона поступает на аналого-цифровой преобразователь (АЦП). Цифровой сигнал с выхода АЦП поступает на блок сегментации. Функционально этот блок является детектором голосовой активности, т. е. устройством, выделяющим в сигнале фрагменты, содержащие речь. На выходе блока сегментации сигнал представлен в виде речевых фрагментов, соответствующих отдельным словам, которые поступают на блок сравнения с эталоном. Выходные данные блока сравнения поступают на устройство принятия решения (УПР).

Описанная схема может иметь разные реализации в зависимости от структуры и принципов функционирования блоков, из которых она состоит. Существует много способов сравнения речевых сигналов. Рассмотрим сравнение динамических спектрограмм (сонограмм) методом нелинейного преобразования масштаба времени (от англ. Dynamic Time Warping, DTW). В русском языке не существует устоявшегося названия рассматриваемого метода. Наиболее близким к англоязычному оригиналу можно считать следующее название – динамическое преобразование времени. Здесь же используется несколько видоизмененное название, более точно характеризующее суть метода.

Спектрограмма (сонограмма, спектрограмма типа «водопад») – двумерный спектрально-временной образ сигнала, получаемый с помощью кратковременного преобразования Фурье (рис. 3). Такая форма представления отражает изменение по времени амплитуд частотных составляющих речевого сигнала и хорошо выражает особенности речи. В задачах распознавания

речевых сигналов спектрограммы более информативны, нежели временное представление.

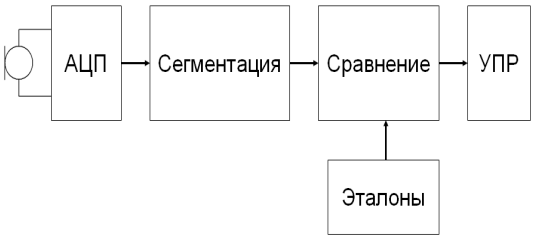

Рис. 2. Схема распознавания изолированных слов, основанная на сравнении с эталоном

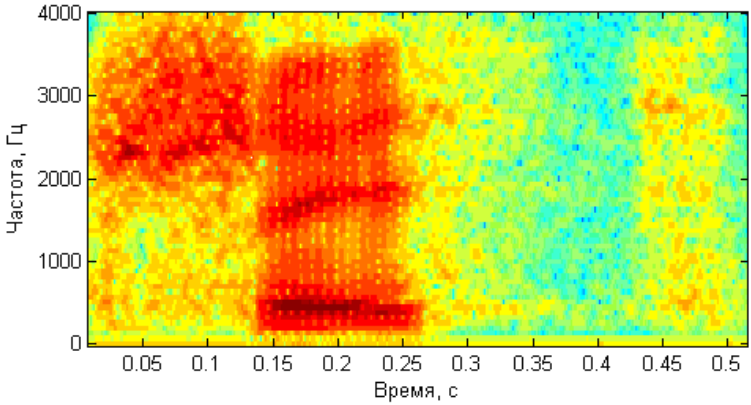

Рис. 3. Спектрограмма, соответствующая слову «шесть»

Однако спектрограммы речевых сигналов, вычисляемые на основе оконного преобразования Фурье, являются не единственным типом признаков, используемых в задачах распознавания. Альтернативным вариантом является использование представления сигнала в виде мел-частотных кепстральных коэффициентов. Алгоритм их вычисления основан на учете психоакустических закономерностей, что позволяет получить более информационно емкое представление речевых сигналов. Применение мел-частотных кепстральных коэффициентов вме-

сто традиционных спектрограмм позволяет повысить точность и помехоустойчивость существующих методов распознавания речевых сигналов.

Различные реализации речевых образов (спектрограмм или наборов мел-частотных кепстральных коэффициентов), относящихся к одному и тому же классу, могут значительно отличаться друг от друга по длительности. Это связано с нестабильностью темпа речи диктора, вызванного влиянием интонации, акцента и т. п. Для корректного сопоставления речевых образов необходимо производить их выравнивание по длине. Выравнивание путём линейного сжатия или растяжения одной реализации слова до величины другой решает задачу лишь частично, так как не учитывается одно важное свойство речевого сигнала – неравномерность его протекания во времени. Это свойство речи выражается в неравномерном изменении длительности звуков слова при изменении длительности слова в целом. Поэтому сопоставление целесообразно выполнять с помощью нелинейного преобразования масштаба времени, основанного на решении оптимизационной задачи поиска наикратчайшего пути между двумя образами методами динамического программирования.

Рассмотрим суть метода нелинейного преобразования масштаба времени применительно к решаемой задаче. Обозначим евклидово расстояние между i-й строкой матрицы входной спектрограммы и j-й строкой матрицы эталона как D_{ij}. Для нахождения строк матрицы входной спектрограммы, наилучшим образом соответствующих строкам матрицы эталона, определяется матрица трансформации C размера $(M \times N)$ по следующим формулам:

$$C(1,1) = D_{11};$$

$$C(i,1) = D_{i1} + C(i-1,1),\ i = 2..M;$$

$$C(1,j) = D_{1j} + C(1,j-1),\ j = 2..N;$$

$$C(i,j) = D_{ij} + \min[C(i-1,j), C(i-1,j-1), C(i,j-1)],\ i = 2..M,\ j = 2..N,$$

где M – количество строк матрицы входного образа; N – количество строк матрицы эталона.

После этого находится оптимальный путь трансформации и стоимость пути. Стоимость пути свидетельствует о близости двух сравниваемых образов и может использоваться для распознавания.

На рисунке 4 ломаной линией соединены элементы матрицы *C*, отвечающие наиболее соответствующим строкам входной спектрограммы и эталона. Вертикальному отрезку соответствует случай, когда несколько строк матрицы эталона соответствуют одной строке матрицы входной спектрограммы; горизонтальному отрезку – случай, когда несколько строк матрицы входной спектрограммы соответствуют одной строке матрицы эталона. Очевидно, что при сравнении абсолютно идентичных образов траектория преобразования будет иметь вид диагонали.

Таким образом, в отличие от алгоритма линейного растяжения, данный алгоритм обеспечивает выравнивание только спектрально подобных фрагментов входной спектрограммы и эталонного образа, что позволяет существенно повысить точность распознавания изолированных слов, представленных в спектральной области.

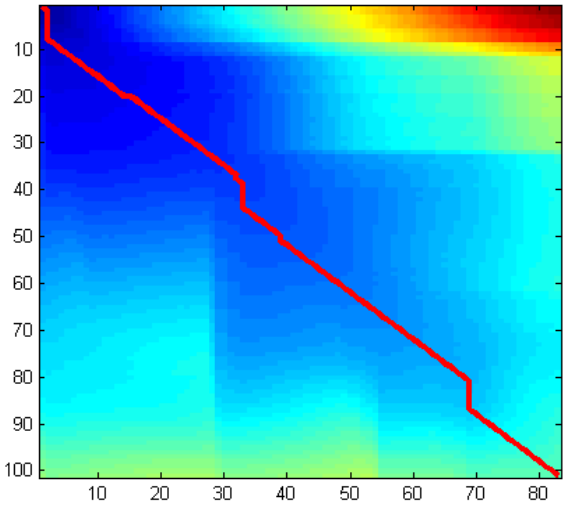

Рис. 4. Графическая интерпретация
метода нелинейного преобразования времени

Практические задания

1. С помощью метода нелинейного преобразования масштаба времени произведите сопоставление идентичных речевых фрагментов. Убедитесь, что в этом случае оптимальным путем трансформации будет диагональ, т. е. отсутствие временного растяжения распознаваемого фрагмента.

2. Проанализируйте, что изменится в случае, если сравнивать аналогичные фрагменты, но один из них предварительно зашумить аддитивным белым гауссовским шумом. Рассмотрите случаи с разными отношениями сигнал/шум (в диапазоне от 0 до 30 дБ).

3. Проанализируйте, что изменится в случае, если сравнивать аналогичные фрагменты, но один из них предварительно усекать (удалять часть отсчетов в начале или конце фрагмента).

4. Используя речевую базу, состоящую из записей речевых команд, произнесенных разными дикторами, произведите следующие количественные исследования:

5. Точность распознавания речевых команд в зависимости от параметров спектрального преобразования: длины окна и коэффициента перекрытия. В исследовании предлагается использовать следующие длины окон: 32, 64, 128, 256, 512 отсчетов – и следующие значения коэффициентов перекрытия межу соседними окнами: 0; 0,25; 0,5; 0;75 (в долях от длины окна). Результат представить в виде графиков. Проанализируйте полученный результат и выберите параметры алгоритма для применения на следующих стадиях исследования;

6. Точность распознавания речевых команд в зависимости от ОСШ (в диапазоне от 0 до 30 дБ с шагом в 10 дБ) для аддитивного белого гауссового шума. Результат представьте в виде графика зависимости процента правильно распознанных команд от ОСШ;

7. Точность распознавания речевых команд в случае неправильной сегментации речевых команд (усечения речевого фрагмента в начале, в конце фразы и с обеих сторон одновременно). Результат представьте в виде графиков зависимости процента

правильно распознанных команд от величины усечения (в долях (от 0 до 0,5 с шагом 0,1) от общей длины речевого фрагмента).

8. Сделайте выводы об устойчивости метода нелинейного преобразования масштаба времени к искажающим воздействиям различного характера (на примере воздействия аддитивного шума и усечения речевых фрагментов при неправильной сегментации).

9. Модифицируйте метод, заменив спектрограммы на наборы мел-частотных кепстральных коэффициентов. Подберите параметры преобразования. Проведите сравнение модифицированного метода с исходным.

Контрольные вопросы

1. В чем состоит задача детектирования голосовой активности?

2. Какие признаки речевых сигналов используются для детектирования голосовой активности?

3. Что понимается под качеством речевого сигнала?

4. В чем отличие между субъективными и объективными методами оценки качества?

5. В чем отличие между эталонными и неэталонными методами оценки качества?

6. Назовите наиболее распространенные объективные методы оценки качества речевых сигналов.

7. В чем состоит идея построения комбинированного показателя качества? Назовите его достоинства и недостатки.

8. Что понимается под разборчивостью речевого сигнала? В чем состоит отличие между разборчивостью и качеством?

9. Как производится субъективная оценка разборчивости речи?

10. Назовите основные группы объективных методов оценки разборчивости.

11. Назовите наиболее распространенные формантные методы оценки разборчивости. В чем состоит общий принцип их работы?

12. В чем состоит основная идея модуляционных методов оценки разборчивости? Какой сигнал используется для измерений?

13. Сформулируйте задачу шумоподавления, а также основные предположения, лежащие в основе методов шумоподавления, работающих в спектральной области.

14. Что такое априорное и апостериорное отношение сигнал/шум?

15. Что такое функция коррекции спектра? Назовите наиболее распространенные функции коррекции спектра.

16. В чем суть идеи использования бинарных масок при решении задачи шумоподавления в спектральной области?

Рекомендуемая литература

1. Рихтер, С. Г. Цифровое радиовещание / С. Г. Рихтер. – М. : Горячая линия-Телеком, 2004. – 352 с.

2. Дидковский, В. С. Акустическая экспертиза каналов речевой коммуникации : монография / В. С. Дидковский, М. В. Дидковская, А. Н. Продеус – Киев : Имекс-ЛТД, 2008. – 420 с.

3. Приоров, А. Л. Обработка и передача мультимедийной информации : учеб. пособие для вузов / А. Л. Приоров, В. В. Хрящев ; Яросл. гос. ун-т им. П. Г. Демидова. – Ярославль : ЯрГУ, 2010. – 187 с.

4. Аграновский, А. В. Теоретические аспекты алгоритмов обработки и классификации речевых сигналов / А. В. Аграновский, Д. А. Леднов. – М. : Радио и связь, 2004. – 164 с.

5. Ахмад, Х. М. Введение в цифровую обработку речевых сигналов : учеб. пособие / Х. М. Ахмад, В. Ф. Жирков ; Владим. гос. ун-т. – Владимир : Изд-во Владим. гос. ун-та, 2008. – 192 с.

Оглавление

Учебное издание

Топников Артем Игоревич

Цифровая обработка речевых сигналов

Практикум

Верстка Е. Б. Половкова
Редактор, корректор Л. Н. Селиванова

Подписано в печать 28.09.18. Формат 60×84 1/16.
Усл. печ. л. 2,32. Уч.-изд. л. 2,0.
Тираж 2 экз. Заказ

Оригинал-макет подготовлен
в редакционно-издательском отделе ЯрГУ.

Ярославский государственный университет
им. П. Г. Демидова.
150003, Ярославль, ул. Советская, 14.