**Обзор представлений звукового сигнала для дальнейшей обработки нейронными сетями в задаче распознавания эмоций**

1. Волновая форма (waveform).

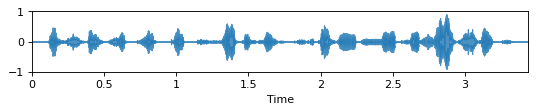
Звук представляет собой непрерывный сигнал — волну с меняющейся амплитудой и частотой. Для обработки такого сигнала, необходимо представить его в виде цифрового. В упрощенном виде, процесс представления сигнала в цифровой форме связан с замером мгновенных значений амплитуды аналогового сигнала (дискретизация по амплитуде) в определенные и постоянные моменты времени (дискретизация по времени). Важнейшими параметрами цифрового сигнала является частота дискретизации – количество замеров, производимых в единицу времени, и глубина цифрового кодирования - количество информации, которое необходимо для кодирования каждого дискретного значения амплитуды. Таким образом, в результате звуковой сигнал можно представить как N-мерный вектор, где N – произведение частоты дискретизации в герцах и длительности звукового сигнала в секундах. При этом каждый элемент этого вектора представляет собой целое число от 1 до 2I, где I – глубина кодирования в битах. Для обработки данного вектора нейронными сетями, значения всех его элементов приводятся в диапазон от -1 до 1 (нормализация). *Программные решения, позволяющие получить такой вектор: библиотеки librosa, scipy, wave.*

Рисунок . Пример волновой формы речевого сигнала.

1. Временно-частотные представления (time-frequency representations).

Способы визуального отображения аудиосигнала через различные частотно-временные представления предлагают богатое представление временной и частотной структуры исходного сигнала. К ним относятся:

* STFT (Short-Time Fourier Transform, кратковременное преобразование Фурье)
* CQT (Constant-Q Transform)
* CWT (Continuous Wavelet Transform)

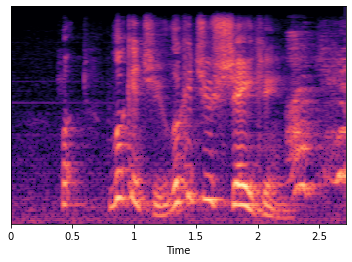
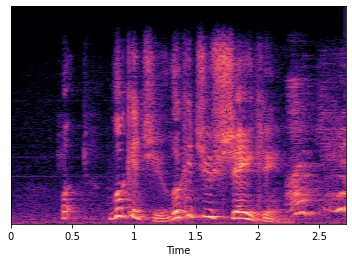
Отображения, получаемые с помощью STFT и CQT называются спектрограммами и представляют собой двумерную матрицу, тогда как отображения, получаемые с помощью CWT – скейлограммами или скалограммами (scalogram) и представляют собой трехмерную матрицу. Спектрограммы дополнительно классифицируются по использованной частотной шкале: линейная, логарифмическая и мел-шкала. Наиболее широко используемой в связанных с речью задачах является логарифмическая и мел-логарифмическая шкала.

Рисунок . Спектрограмма речевого сигнала, логарифмическая (слева) и мел-логарифмическая (справа) шкала.

Временно-частотные отображения являются гораздо более информативными по сравнению с волновой формой представлениями звукового сигнала, поскольку добавляется дополнительное измерение: по горизонтальной оси откладывается время, по вертикальной – частота, третьим измерением является интенсивность пикселя. *Существует множество программных решений, позволяющих получать такие отображения: Matlab, библиотеки librosa, scipy, pytorch (с вер. 1.7).*

1. Признаковые представления.

До 2008 года именно данные представления наиболее широко использовались в машинном обучении. Способ данного представления заключается в извлечении ряда признаков из звукового сигнала и обработке этого ряда. Признаки, извлекаемые из звукового сигнала, делятся на низкоуровневые дескрипторы (low-level-descriptors, LLD) и функциональные признаки. Низкоуровневые дескрипторы включают в себя просодические (высота тона, громкость, энергия, тембр, продолжительность пауз и др.) и спектральные (фундаментальная частота, часоты основных формант, мел-кепстральные частотные коэффициенты (MFCC), кепстральные коэффициенты линейного предсказания (LPCС) и др.) характеристики, а так же их производные по времени. Функциональные признаки включают в себя статистические показатели низкоуровневых дескрипторов (минимум, максимум, различные процентили, zero-crossing-rate).

*Для извлечения и обработки этих признаков широко используется инструмент OpenSMILE*, позволяющий извлекать широкий спектр параметров звукового сигнала, а также применять различные функции к этим параметрам. Благодаря этому, появились готовые наборы параметров, которые могут быть легко извлечены с помощью этого инструмента. К этим наборам относятся как очень большие, как, например, стандартные наборы признаков для конференций INTERSPEECH, содержащие более 5000 элементов (СomPaRE-2014), так и попытка унифицировать признаковое пространство для задач аффективных вычислений (GeMAPS и eGeMAPS).

GeMAPS и eGeMAPS стоит выделить отдельно, так как это относительно компактный набор (88 параметров, среди которых: статистические функции от основной частоты и амплитуды, мел-кепстральные коэффициенты (MFCC) 1-4, приближенная оценка количества слов в секунду), тем не менее отлично показывающий себя в задачах распознавания эмоций.