УДК 0004.934

ОЦЕНКА КАЧЕСТВА РЕЧЕВОГО СИГНАЛА С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Волкова М.В. (Университет ИТМО), **Кабаров В.И.** (Университет ИТМО),
Астапов С.С. (Университет ИТМО)
Научный руководитель – к.т.н., Новоселов С.А.
(Университет ИТМО)

Аннотация. В работе рассматривается способ обучения нейронных сетей оценивать такие акустические характеристики речевых сигналов, как отношение сигнал/шум, время реверберации, предсказывать тип шума, присутствующего в записи, и давать общую (интегральную) оценку качества как функцию от указанных оценок на целом речевом сигнале или его фрагменте. При этом оценка качества производится только по данному входному сигналу, без необходимости сравнения с эталонным (неискаженным) сигналом, а для оценки времени реверберации не требуется знать импульсную характеристику помещения, в котором была сделана запись речи.

Развитие речевых технологий в различных областях, в том числе голосовой биометрии, показало зависимость таких систем от качества речевого сигнала. Система, обученная на данных из определённого источника, домена, часто показывает снижение качества на данных в другом домене. Для адаптации системы к целевому домену может быть полезно знать его акустические условия: уровень и тип шума, а также время реверберации.

В данной работе предлагается способ оценки качества речевого сигнала без необходимости сравнения с эталонным (неискажённым) сигналом, что в литературе часто называется «слепой оценкой» (blind estimation). Для этого был разработан способ обучения нейронной сети предсказывать отношение сигнал/шум (SNR), уровень реверберации (RT60) и тип шума, а также давать обобщённую оценку качества. Оценка проводится на интервалах в 2 секунды, либо даётся усреднённая оценка на всём файле целиком.

Оценка качества речевого сигнала может использоваться в различных приложениях речевой обработки, например, в задаче автоматического выбора наилучшего микрофона в многомикрофонной системе записи звуковых сигналов. В случае голосовой биометрии она может использоваться для определения наиболее качественных сегментов речи в записях, выполненных в различных акустических условиях, для построения голосовой модели диктора по выбранным фрагментам.