УДК 519.86

ОТДЕЛЕНИЕ ГОЛОСА ОТ МУЗЫКИ ИЗ АУДИОЗАПИСЕЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ

В. С. БУТОМА, А. Н. ДУХОВНИК
Научный руководитель В. Н. КОЗУБ
УО «Белорусский государственный университет информатики и радиоэлектроники»
Минск, Беларусь

В настоящее время задача разделения голоса и музыки не является решенной. Одним из возможных применений является использование системы разделения голоса от фонограммы в караоке-системах и мобильных приложениях, которые в настоящее время становятся крайне популярными. Существуют решения с различной степенью качества, например, встроенные в стандартные утилиты для обработки звука Audacity и Soundmane и т. д. В данных утилитах используется следующие методы: вырезание голосовых частот, вырезание центра спектрограммы.

Особенный интерес эта задача представляет для решения ее с помощью нейронный сетей, в том числе с помощью архитектур со сверточными слоями.

Предлагается использовать нейронную сеть, которая будет представлять собой сверточный автоенкодер со связями между последовательными слоями и связями между ранними и более поздними слоями. Для представления звука как спектрограммы предлагается использовать оконное преобразование Фурье.

Алгоритм обучения нейронной сети:

– аудиофайл раскладывается в спектрограмму (трехмерный массив), которая затем отправляется на вход нейронной сети;

– спектрограмма, полученная на выходе из нейронной сети, сравнивается с эталонной спектрограммой голоса, при этом разность между ними используется для обновления весов в сети методом обратного распространения ошибки;

– полученную спектрограмму можно использовать для создания аудиофайла методом обратного преобразования Фурье.

Для получения набора данных при обучении можно применить открытые базы данных записей голоса, записанного в акустических студиях, в которых нет фонограммы. Аналогично можно воспользоваться данными фонограмм и минусов для получения набора данных о музыке без голоса. Далее простым наложением музыки и голоса получаем набор данных, который будет поступать на вход нейронной сети, а исходные записи с «чистым» голосом используем как эталон на выходе из сети.