МЕТОДОЛОГИЯ ОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ ИЗОЛЯЦИИ ГОЛОСА ИЗ КОМПОЗИЦИИ

К.В. Рубанов
И.А. Мурашко*, научный руководитель, д-р техн. наук, профессор*
Гомельский государственный технический университет имени П.О. Сухого
г. Гомель

Старые аудиозаписи часто имеют на фоне посторонние шумы и низкое качество инструментального аудиоряда, а получение чистого вокального ряда таких песен может оказаться невозможным. Решением этой проблемы может стать изоляция из аудиозаписи вокального ряда.

Для решения подобной задачи существует множество методов, основанных на знании о физических качествах голоса и нейронных сетях. Множество готовых решений являются коммерческой тайной либо не предоставляют достаточного уровня качества обработки аудиозаписей.

Основная цель данной научной работы заключается в выработке методологии и подготовке программного продукта для изоляции голоса из музыкальной композиции.

Для обучения любой нейронной сети необходимо подготовить обучающую выборку, данные в которой послужат основой для принятия последующих решений обученной моделью. При обучении с учителем исходный набор данных должен включать в себя массив групп признаков для каждого этапа обучения и маску для каждой такой группы.

При обучении нейронной сети изоляции вокального аудиоряда в музыкальной композиции необходим список аудиозаписей, содержащих как вокальный, так и инструментальный ряд для разбиения ее на фрагменты с признаками, и списком тех же аудиозаписей с исключенным из них инструментальным или любым другим аудиорядом, не относящимся к человеческому вокалу, – такая аудиозапись будет выполнять роль маски.

Наименее трудозатратный способ составления обучающей выборки: создание собственных аудиозаписей путем соединения любого вокального и любого музыкального аудиорядов в одну композицию. При таком подходе в выборке будет присутствовать наиболее точная маска для композиции. Недостаток подхода: полученные данные в недостаточной мере описывают реальные композиции и не дают ожидаемого результата при выделении вокала из фрагментов с инструментами, близкими по частоте к человеческому голосу.

Полученные аудиозаписи разбиваются на фреймы, каждый из которых состоит из 513 сэмплов. Фреймы необходимо собрать во фрагменты, каждый из которых содержит 25 фреймов. Технология составления фрагментов состоит в выделении STFT окна [1], состоящего из 25 фреймов, каждый фрагмент образуется смещением STFT окна на один фрейм, начиная с нулевого и заканчивая n-25-ым фреймом. Таким образом, из n-фреймов можно получить n-24 фрагмента, поскольку значимыми для каждого фрагмента являются только средние фреймы. В качестве ответа сети следует использовать фрейм из аудиозаписи-маски, соответствующий временному интервалу среднего фрейма исследуемого фрагмента.

Модель нейронной сети состоит из 4 сверточных слоев [2] и следующими за ними слоями субдистретизации и слоями активации. И многослойного персептрона, содержащего 3 слоя. Последний слой возвращает ответ в виде одномерного тензора с 513 элементами.

Полученный ответ нейронной сети необходимо привести к виду бинарной маски, применяя пороговую функцию к каждому элементу одномерного тензора. Поэлементное умножение маски на аудиозапись позволит обратить в ноль сэмплы, не относящиеся к вокальному аудиоряду.

Нейронная сеть была обучена на 140 фрагментах с 3 эпохами обучения. Полученный результат обработки композиции группы Король и Шут «Лесник» можно визуализировать, приведя для сравнения исходную аудиозапись (рис.).

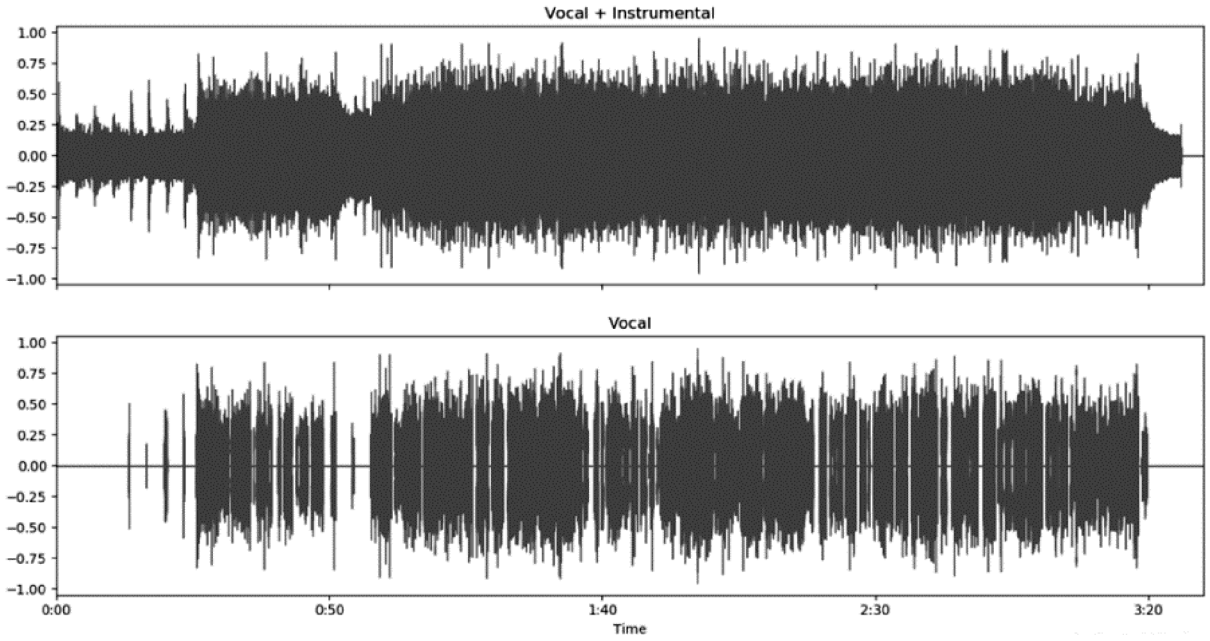

Рис. Результат обработки аудиозаписи

Результатом работы является нейронная сеть, способная изолировать вокальный аудиоряд из композиции. Качество обработки полностью зависит от архитектуры сети и обучающей выборки. Полученная сеть способна отделять вокальный аудиоряд от инструментального, но на итоговой аудиозаписи присутствуют артефакты и смесь вокала с инструментами, что можно исправить увеличением объема и качества обучающей выборки.

1. Адаптивное разделение источников звука в режиме реального времени: заявка 15 / 434, 419 Соединенные Штаты Америки / Р. Пилл; заявитель Ред Пилл; патент проверенный Коретский и др. – 9,842,609 B2; заявл. 12.01.17; опубл. 17.08.17; приоритет 16.02.16, N US201662295497P(США).

2. Реализация метода разделения речи с помощью глубоких нейронных сетей в режиме реального времени: заявка 14/536,114 Соединенные Штаты Америки / Ш. Кампбелл; заявитель Шэнон Кампбелл; патент. проверенный Вонг и др. – N 2017/0061978; заявл. 07.04.14; опубл. 02.03.17; приоритет 07.11.14, N 14/536,114 (США).