УДК 004.42 **Информационные технологии**

Яковлев Александр Николаевич, *студент КФ МГТУ им. Н.Э. Баумана*
Бурмистров Александр Викторович, *ассистент кафедры «Защита информации» КФ МГТУ им. Н.Э. Баумана*
Крысин Иван Александрович, *ассистент кафедры «Информационные системы и сети» КФ МГТУ им. Н.Э. Баумана*

СИСТЕМА РАСПОЗНАВАНИЯ ДИКТОРА НА ОСНОВЕ АЛГОРИТМА CREPE

Аннотация: В статье раскрывается понятие частоты основного тона, приводится описание нейросетевого алгоритма CREPE. Приводится описание архитектуры нейронной сети, используемой для распознавания дикторов. Описано тестирование системы на основе данных датасета VoxCeleb.

Ключевые слова: Распознавание диктора, сверточная нейронная сеть, функция основного тона, pitch-tracking, CREPE, ReLU.

Annotation: The paper reveals the concept of the frequency of the main tone, provides a description of the neural network algorithm CREPE. The description of the neural network architecture used to recognize speakers is given. The testing of the system based on the VoxCeleb dataset is described.

Keywords: Speaker recognition, convolutional neural network, pitch function, pitch-tracking, CREPE, ReLU.

Введение

Определение личности на основе аудиоматериала является актуальной задачей, находящей широкое применение в различных сферах, таких как голосоведение,

криминалистика, методы идентификации пользователя, распознования речи и многих схожих задачах [5]. Особую актуальность данная проблема приобретает в свете постоянного роста функционала различных голосовых ассистентов (таких как Siri, Cortana, Google Assistant, Alexa и им подобных). Для такого рода помощников (все чаще встраиваемых в системы голосового управления различными смарт-устройствами), необходимо не только успешно распознавать речь (в рамках отдельных слов, словосочетаний и предложений), но и иметь возможность соответствующим образом определять личность пользователя, во избежание несанкционированного доступа.

Одной из ключевых характеристик голосового сигнала, является так называемая частота основного тона (англ. Fundamental frequency или fundamental, обозначается как F0). Особенностью данного параметра для голоса является его уникальность – с музыкальной точки зрения, частота основного тона является образующей для всех остальных звуков натурального звукоряда, а частоты кратные F0 называются гармониками. Для человеческой речи, частота основного тона – это частота, при которой открываются и закрываются голосовые связки, а ее повышение воспринимается слушателем как повышение высоты звука. Таким образом идентификацию диктора можно свести к задаче определения частоты основного тона [1].

В настоящее время задача определения частоты основного тона аудио и речи (т.е. задача pitch-tracking’a) является одной из ключевых для сферы обработки звука и находит множество применений, таких как разбор музыкального материала для последующего его анализа, автоматическая транскрипция музыки, эффективное сжатие аудиофайлов, а также обучение игре на музыкальных инструментах. Стоит отметить, что в настоящее время существует множество инструментов, позволяющих решить задачу определения частоты основного тона. Данный инструментарий постоянно обновляется и дополняется новыми решениями в этой области.

Целью данного исследования является изучение процесса идентификации пользователя по его голосовому аудиоматериалу в рамках алгоритма определения частоты основного тона на основе сверточной нейронной сети. Алгоритм, имеющий название CREPE – A Convolutional REpresentation for Pitch Estimation, был представлен 17 февраля 2018 года и является новым решением в области задач, направленных на определение частоты основного тона в монофоническом аудиоматериале.

Алгоритм CREPE

Алгоритм CREPE предоставляет возможность определения частоты основного тона на основе фактических данных. Благодаря этому, данный метод дает более точные результаты для процесса определения по сравнению с уже существующими алгоритмами на базе эвристики (таких как SWIPE и pYIN [3]). В основе архитектуры алгоритма находится сверточная нейронная сеть, оперирующая непосредственно аудио сигналом во временной области. Диаграмма архитектуры трекера высоты звука, основанного на рассматриваемом алгоритме представлена на рис. 1

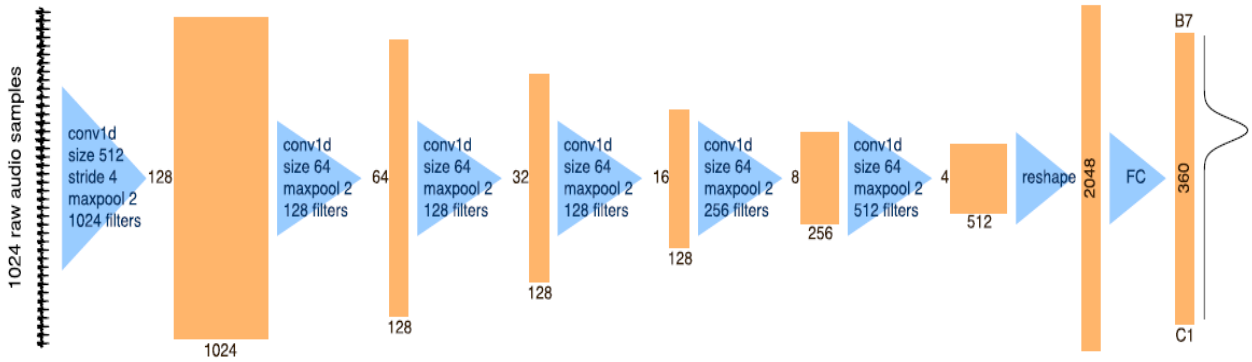

Рис. 1. Архитектура трекера высоты звука, основанного на алгоритме CREPE

Трекер высоты звука, спроектированный на основе алгоритма CREPE имеет следующий принцип работы: в качестве входных данных взяты 1024 выдержки из аудио сигнала во временной области с частотой дискретизации 16 кГц и обрабатываемые посредстовм шести сверточных слоев (и 6 слоев

субдискретизации соответственно) [2]. Выходными данными является скрытое представление в виде вектора размерностью 2048. Это представление тесно соединено с выходным слоем с активирующей функцией-сигмоидой, представляющим собой 360-размерный вектор. Каждый из 360 элементов выходного вектора соответствует определенному значению высоты звука, выражаемой в центах. Цент – единица частотного интервала, равная сотой части полутона или 1/1200 части октавы (поскольку в октаве 12 полутонов) и является функцией от частоты:

$$¢(f) = 1200 \times \log_2 \frac{f}{f_{ref}}$$

где $f_{ref} = 10$ кГц, что дает шкалу высот звука, в которой 100 центов равняется одному полутону. Таким образом, данная шкала покрывает диапазон звуков от C1 до B7 (от ноты До контроктавы до ноты Си четвертой октавы) с интервалами в 20 центов, в диапазоне частот от 32,70 Гц до 1975,5 Гц.

Для решения задачи идентификации говорящего на основе его аудиоматериала становится возможным применить подобный механизм. В рамках данного исследования был создан трекер частоты основного тона на основе сверточной нейронной сети (СНС), представленный в табл. 1

Таблица 1. Архитектура СНС

Слой	**Кол-во фильтров**	**Размер ядра/выборки**	**Размер выходных данных**
Conv1d_1	1024	512	256x1024
MaxPool1d_1	-	2	128x1024
Conv1d_2	128	64	128x128
MaxPool1d_2	-	2	64x128
Conv1d_3	128	64	64x128
MaxPool1d_3	-	2	32x128
Conv1d_4	128	64	32x128
MaxPool1d_4	-	2	16x128

Conv1d_5	256	64	16x256
MaxPool1d_5	-	2	8x256
Conv1d_6	512	64	8x512
MaxPool1d_6	-	2	4x512
Dropout	-	-	4x512
Reshape	-	-	2048
Dense	-	-	1251

Одним из важных этапов разработки нейронной сети является выбор функции активации. Именно от нее будет зависеть функционал сети и метод обучения. В настоящее время чаще всего используются следующие виды функции активации:

- линейный выпрямитель (ReLU – Rectified Linear Unit) или её различные вариации;

- сигмоида;

- гиперболический тангенс.

В данной сверточной нейронной сети использована функция активации ReLU, так как она обладает высокой производительностью, а также простотой понимания и использования.

Стоит отметить, что для корректной работы алгоритма идентификации (а также для возможности грамотной оценки производительности такого подхода) необходим аудиоматериал с очень качественной и точной аннотацией [4]. Поэтому в качестве рабочего набора данных была взята выдержка из датасета VoxCeleb, созданного в 2017 году силами Оксфордского университета и представляющего собой сборник аудиоматериалов из 153516 высказываний 1251 различных знаменитостей, записанных с частотой дискретизации 16 кГц. Выдержка из датасета представляет собой 500 аудиозаписей, обучение сети велось с разделением данных в отношении 80:20.

Результаты работы сети представлены в таблице 2.

Таблица 2. Результат работы сверточной нейронной сети

Значение функции потерь	0.09
Точность работы	82%

Выводы

Относительно низкая точность работы сверточной нейронной сети в данном случае обуславливается малым количеством аудиоматериала в обрабатываемом датасете (400 аудиозаписей для тренировочного набора и 100 для набора тестирования). Однако, даже при таких значениях можно говорить о возможности идентификации говорящего на основе его записанного аудиоматериала при помощи сверточной нейронной сети.

Библиографический список:

1. Alain de Cheveigne, Hideki Kawahara, YIN, a fundamental frequency estimator for speech and music, Ircam-CNRS, Wakayama University, 2002. URL: http://recherche.ircam.fr/equipes/pcm/cheveign/ps/2002_JASA_YIN_proof.pdf (дата обращения 15.06.2019).

2. Jong Wook Kim, Justin Salamon, Peter Li, Juan Pablo Bello, CREPE: a convolutional representation for pitch estimation, Music and Audio Research Laboratory, New York University, Center for Urban Science and Progress, New York University, 2018. URL: https://arxiv.org/pdf/1802.06182.pdf (дата обращения 18.06.2019).

3. Pitch-tracking, или определение частоты основного тона в речи на примерах алгоритмов Praat, YAAPT и YIN / [Электронный ресурс]. https://habr.com/company/neurodatalab/blog/416441/ (дата обращения 23.06.2019).

4. Бурмистров А.В., Гришунов С.С., Молчанов А.Н. К вопросу об эффективности систем верификации пользователей по голосу // Электронный

журнал: наука, техника и образование. 2017. №1 (10). С. 16-20. URL: http://nto-journal.ru/uploads/articles/c6431af540bead7a85f948ee184dc422.pdf (дата обращения 07.07.2019).

5. Бурмистров А.В., Гришунов С.С., Молчанов А.В. Математические методы классификации дикторов // Вопросы радиоэлектроники. 2016. № 10. С. 13-17 (дата обращения 04.07.2019).