РАСПОЗНАВАНИЕ РЕЧИ НА ОСНОВЕ АНАЛИЗА АРТИКУЛЯЦИИ ГОВОРЯЩЕГО ПО ВИДЕОПОТОКУ С ИСПОЛЬЗОВАНИЕМ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ ГЛУБОКОГО ОБУЧЕНИЯ

К. Д. Левшун

Белорусский государственный университет, г. Минск;
klmail98@gmail.com; науч. рук. – И. А. Адуцкевич

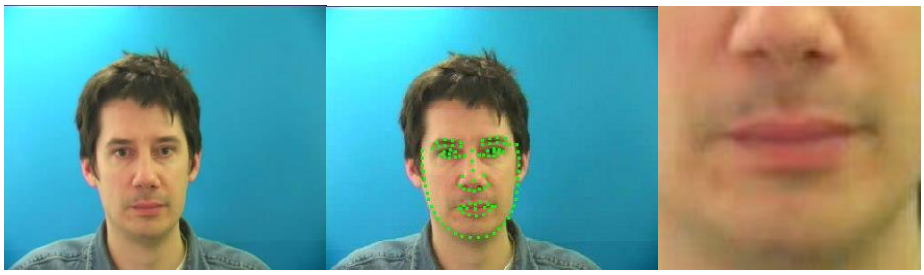
В статье приведено описание реализованной системы распознавания речи в видеопотоке на основе нейросетевых алгоритмов компьютерного зрения. В ходе работы произведена обработка данных, обучение и тестирование нейросетевых моделей на наборе данных Gridcorpus. Реализованная система может применяться для реабилитации людей, потерявших возможность говорить, а также для распознавания речи в условиях повышенного уровня шума.

Ключевые слова: распознавание речи; нейронные сети; LipNet.

Распознавание речи по губам является достаточно сложной задачей для людей. По имеющимся исследованиям, слабослышащие определяют только 17 $\pm$ 12 % из ограниченного набора 30-ти односложных слов и 21 $\pm$ 11 % из 30-ти многосложных слов [1]. Поэтому автоматическое распознавание имеет большой практический потенциал для применения в зашумленных помещениях, идентификации личности, а также в медицинских целях.

Для исследования использовался набор данных GRID [2], состоящий из видеозаписи 1000 предложений на английском языке, произнесенных 34-мя докладчиками. При этом некоторые видеозаписи повреждены, после фильтрации остается 32904 видеозаписи длительностью 3 секунды каждая. Каждая видеозапись делится на 75 кадров. Данные разделяются на тренировочную (26304 записи) и отложенную (6600 видеозаписей) выборки.

Для выделения необходимого участка лица используется детектор ключевых точек из библиотеки dlib [3], и получаемые точки соответствуют определенным участкам лица. Так как в нашей работе используются только артикуляционные движения губ, это позволяет нам использовать лишь участок лица, соответствующий губам, а именно точкам 48–67.



а) б) в)

Рис. 1. Выделение участка губ:
(а) – исходное изображение, (б) – получение ключевых точек,
(в) – результат преобразования

Для перевода потока кадров в текст используется нейронная сеть архитектуры LipNet [4].

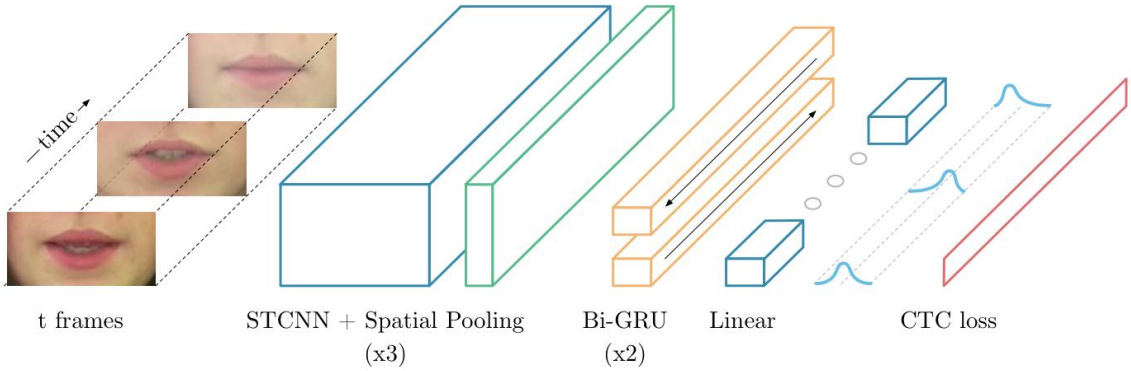


Рис. 2. Архитектура нейронной сети LipNet

В нейронной сети LipNet используются слои пространственно-временных сверток (STCNN) [5], управляемые рекуррентные блоки (GRU) [6], а также полносвязный слой.

Обучение нейронной сети производилось при помощи метода оптимизации Adam [7] с размером шага 10^{-4}, в качестве функции потерь используется CTC loss [8], а для декодирования выхода нейронной сети в текст - CTC beam decoder [9].

Для оценки качества предсказания (таблица) использовалась метрика WER (Word Error Rate), вычисляемая по формуле:

$$WER = \frac{S + D + I}{N},$$

где S – количество замен, D – количество удалений, I – количество вставок, N – количество слов в предложении.

Средняя оценка WER для предсказаний полученной модели – 22,7 %, что значительно меньше, чем оценка предсказаний речи людьми – 47,7 % [4].

Таким образом, приведенная система позволяет производить распознавание речи по видеопотоку качественнее, чем это удается людям.

Таблица

Пример результатов распознавания

Оригинальный текст	Предсказанный текст	WER
place red by m seven please	place red by m seven please	0,0
place green at t seven now	place green at d seven now	0,1667
place green in t six again	place green at d six again	0,3333
set blue by q seven now	set blue by u seven now	0,1667
place red in z six soon	place red i c six soon	0,3333
place white by n two soon	place red by h two soon	0,1333

Реализованная система может быть использована людьми, потерявшими способность воспроизводить звуки речи ввиду перенесенных болезней, травм и медицинских манипуляций, таких как, например, трахеостомия (введение в трахею трубки при непроходимости дыхательных путей) или удаление гортани для борьбы с раком. Также, такая система может быть использована для распознавания речи в условиях, где записать голос человека не представляется возможным, например при повышенном уровне шума, удаленности человека от установленного прибора. В дальнейшем планируется проведение экспериментов с другими нейросетевыми моделями, а также сбор собственных данных для построения системы распознавания речи на русском языке.

Библиографические ссылки

1. *R. D. Easton, M. Basala,* Perceptual dominance during lipreading // Perception & Psychophysics, 1982, pp 562–570
2. The GRID audiovisual sentence corpus [Электронный ресурс] URL: http://spandh.dcs.shef.ac.uk/gridcorpus/_(дата обращения: 24.03.2019)
3. Dlib C++ Library [Электронный ресурс] URL: http://dlib.net/ (дата обращения: 24.03.2019)
4. *Yannis M Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas*. LipNet: End-to-end sentence-level lipreading // GPU Technology Conference, 2017.
5. *A. Krizhevsky, I. Sutskever, G. E. Hinton*. Imagenet classification with deep convolutional neural networks // In Advances in neural information processing systems, 2012, pp. 1097–1105.
6. . *Chung, C. Gulcehre, K. Cho, and Y. Bengio*. Empirical evaluation of gated recurrent neural networks on sequence modeling // arXiv preprint arXiv:1412.3555, 2014.
7. *D. Kingma, J. Ba*. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980, 2014.
8. *A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber*. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks // ICML, 2006, pp. 369–376.
9. *D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, et al*. Deep Speech 2: End-to-end speech recognition in English and Mandarin // arXiv preprint arXiv:1512.02595, 2015.