UDC 004.8

Дмитро Смішний,
Олексій Алещенко

РОЗПІЗНАВАННЯ МОВИ З ВИКОРИСТАННЯМ ТЕХНОЛОГІЇ ГЛИБОКОГО НАВЧАННЯ

SPEECH RECOGNITION WITH DEEP LEARNING TECHNOLOGY USING

В даній статті розглядається проблема розпізнавання мови користувача за допомогою технології глибокого навчання. Для вирішення поставленої задачі було також застосовано підхід з рекуретними нейронними мережами та перетворенням Фур'є у ролі математичної моделі. Текст диктується через мікрофон BehringerC-1U. Словом для навчання мережі виступає «Привіт».

Ключові слова: рекурентні нейронні мережі, глибоке навчання, перетворення Фур'є, дискретизація.

Рис.: 4. Бібл.: 5.

This article examines the problem of recognizing speech using deep learning technology. I was also used an approach with recurrent neural networks and Fourier transformation as a mathematical model to solve the problem. The text is dictated by the Behringer C-1U microphone. The word for training the network is Ukrainian hello – "pryvit".

Key words: recurrent neural networks, deep learning, Fourier transformation, sampling.

Fig.: 4.Bibl.: 5.

Актуальність теми дослідження.В нашому інформаційному світі все частіше постає проблема комунікації з тими, чи іншими пристроями одночасно. Оскільки для кожного пристрою потрібно придумувати власний користувацький інтерфейс. Куди як зручніше, для користувача, використовувати голосові команди для керування різноманітними девайсами. Адже це позбавляє додаткових витрат для розробки того ж користувацького інтерфейсу. Та й інструкції до таких пристроїв будуть набагато зрозуміліші і простіші. Ну а використання глибокого навчання, в даному випадку, здешевлює і уніфікує розробку технології голосового керування для різних мов і країн. Дана стаття присвячена проблемі голосового керування саме українською, адже, на жаль, таких систем саме для нашої мови поки дуже мало.

Постановка проблеми. Практично повна відсутність систем розпізнавання голосу для українського сегменту користувачів Інтернету, створених на основі глибокого навчання. А рішення, які вже готові, недосконалі.

Аналіз останніх досліджень і публікацій. За останні роки можна побачити значне пожвавлення розвитку систем розпізнавання голосу. Серед найвідоміших прикладів це AmazonAlexa, GoogleNow, Siri, Алиса від Yandex. Всі вони засновані на нейронних мережах. Однак, варто зауважити, що вони або не підтримують українську, або підтримка даної мови вкрай обмежена. Тому в статті розглядається специфіка розпізнавання тексту українською.

Виділення недосліджених частин загальної проблеми. В даній статті освітлена проблематика розпізнавання голосу, якщо людина говорить українсь-

кою. При цьому, використовується технологія глибокого навчання. Дослідження сфокусовані на генерації тексту на основі голосу, як такого, та застосування даної технології по відношенню до української мови.

Постановка завдання. Завданням даної статті є створення моделі, що буде працювати з українськими словами (в даному випадку, слово «привіт»), розпізнавати їх, та видавати правдивий текст відповідно до надиктованого.

Викладення основного матеріалу.Голос, як і будь-який звук, представляє собою неперервну хвилю з різною амплітудою в різні моменти часу. Оскільки, сучасний комп'ютер - машина цифрова, а не аналогова, то йому для розуміння інформації, що передається голосом, потрібне представлення останньої у вигляді цифрового сигналу. Для цього використовують дискретизацію. В загальному, це технологія, яка дозволяє отримувати данні в певний момент часу з певною частотою [1]. Для людської мови достатньо частоти дискретизації в 16 кГц (16000 разів за секунду отримуємо значення амплітуди). Тепер оцифруємо наше«Привіт» 16000 разів на секунду. Перші 100 точок наведені на рисунку:

```
[-1274, -1252, -1160, -986, -792, -692, -614, -429, -286, -134, -57, -41, -169, -456, -450, -541, -761, -1067, -1231, -1047, -952, -645, -489, -448
, -397, -212, 193, 114, -17, -110, 128, 261, 198, 390, 461, 772, 948, 1451, 1974, 2624, 3793, 4968, 5939, 6057, 6581, 7302, 7640, 7223, 6119, 5461,
 4820, 4353, 3611, 2740, 2004, 1349, 1178, 1085, 901, 301, -262, -499, -488, -707, -1406, -1997, -2377, -2494, -2605, -2675, -2627, -2500, -2148, -
1648, -970, -364, 13, 260, 494, 788, 1011, 938, 717, 507, 323, 324, 325, 350, 103, -113, 64, 176, 93, -249, -461, -606, -909, -1159, -1307, -1544]
```

Рис.1. Кожне число показує амплітуду звукової хвилі
як 1/16000-ої від секундного інтервалу

Можна припустити, що внаслідок такої дискретизації втрачається частина інформації. Тому, для більш точного відображення звуку скористаємося теоремою відліків Віттакера — Найквіста — Котельникова — Шеннона:

$$x(t) = \sum_{k=-\infty}^{\infty} x(k\Delta) sinc\left[\frac{\pi}{\Delta}(t - k\Delta)\right], \text{де } sinc(x) = \frac{\sin(x)}{x}, \quad (1)$$

$$\text{а інтервал дискретизації } 0 < \Delta \leq \frac{1}{2f_c}.$$

Таким чином, ми отримаємо результат, зображений на рисунку 2.

Для полегшення роботи нейронної мережі, ми розділили запис на 20 мілісекундні інтервали, і навчатимемо мережу на них.

Ми розкладаємо звукову хвилю на простіші звукові хвилі, з яких вона складається. Маючи окремі звукові хвилі, ми додаємо потужність звуку для кожної з них.

```
[-1274, -1252, -1160, -986, -792, -692, -614, -429, -286, -134, -57, -41, -169, -456, -450, -541, -761, -1067, -1231, -1047, -952, -645, -489, -448
, -397, -212, 193, 114, -17, -110, 128, 261, 198, 390, 461, 772, 948, 1451, 1974, 2624, 3793, 4968, 5939, 6057, 6581, 7302, 7640, 7223, 6119, 5461,
 4820, 4353, 3611, 2740, 2004, 1349, 1178, 1085, 901, 301, -262, -499, -488, -707, -1406, -1997, -2377, -2494, -2605, -2675, -2627, -2500, -2148, -
1648, -970, -364, 13, 260, 494, 788, 1011, 938, 717, 507, 323, 324, 325, 350, 103, -113, 64, 176, 93, -249, -461, -606, -909, -1159, -1307, -1544,
-1815, -1725, -1341, -971, -959, -723, -261, 51, 210, 142, 152, -92, -345, -439, -529, -710, -907, -887, -693, -403, -180, -14, -12, 29, 89, -47, -
398, -896, -1262, -1610, -1862, -2021, -2077, -2105, -2023, -1697, -1360, -1150, -1148, -1091, -1013, -1018, -1126, -1255, -1270, -1266, -1174, -10
03, -707, -468, -300, -116, 92, 224, 72, -150, -336, -541, -820, -1178, -1289, -1345, -1385, -1365, -1223, -1004, -839, -734, -481, -396, -580, -52
7, -531, -376, -458, -581, -254, -277, 50, 331, 531, 641, 416, 697, 810, 812, 759, 739, 888, 1008, 1977, 3145, 4219, 4454, 4521, 5691, 6563, 6909,
6117, 5244, 4951, 4462, 4124, 3435, 2671, 1847, 1370, 1591, 1900, 1586, 713, 341, 462, 673, 60, -938, -1664, -2185, -2527, -2967, -3253, -3636, -38
59, -3723, -3134, -2380, -2032, -1831, -1457, -804, -241, -51, -113, -136, -122, -158, -147, -114, -181, -338, -266, 131, 418, 471, 651, 994, 1295,
 1267, 1197, 1291, 1110, 793, 514, 370, 174, -90, -139, 104, 334, 407, 524, 771, 1106, 1087, 878, 703, 591, 471, 91, -199, -357, -454, -561, -605,
-552, -512, -575, -669, -672, -763, -1022, -1435, -1791, -1999, -2242, -2563, -2853, -2893, -2740, -2625, -2556, -2385, -2138, -1936, -1803, -1649,
 -1495, -1460, -1446, -1345, -1177, -1088, -1072, -1003, -856, -719, -621, -585, -613, -634, -638, -636, -683, -819, -946, -1012, -964, -836, -762,
 -788]
```

Рис. 2. Двадцяти мілісекундний фрагмент дискретизованого запису слова «привіт»

Це можна зробити за допомогою перетворення Фур'є. Ми розкладаємо звукову хвилю на простіші звукові хвилі, з яких вона складається. Маючи окремі звукові хвилі, ми додаємо потужність звуку для кожної з них [3].

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ix\omega} dx. \quad (2)$$

В результаті, отримаємо оцінку важливості кожного частотного діапазону, від низьких частот до високих:

Рис. 3. Представлення кількості енергії в 50 Гц частотних відрізках

Тепер можна скласти спектрограму, що ще полегшить роботу нейронної мережі, адже їй буде простіше знаходити шаблони в таких даних, ніж в простих записах звуку:

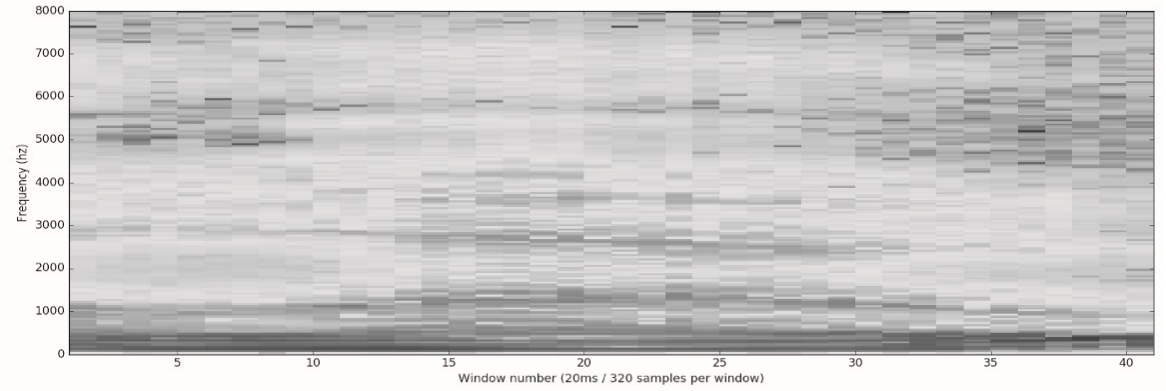

Рис. 4. Повна спектрограма аудіозапису слова «привіт»

Використання ж рекурентних нейронних мереж значно пришвидшить навчання самої мережі. Рекурентні нейронні мережі – це клас штучних нейронних мереж, у якому, з'єднання між вузлами – орієнтований цикл [4]. Це, в свою чергу, створює внутрішній стан мережі, що дозволяє їй проявляти динамічну поведінку в часі. А це означає, що дана мережа на кожному наступному кроці враховує результати попередніх кроків [5]. Звідси, кожна буква, визначена мережею, повинна впливати на ймовірну наступну букву. Наприклад, якщо ми сказали «прив», то швидше за все, далі скажімо «іт», щоб закінчити слово

«привіт». Ймовірність, що ми скажемо щось невимовне, наприклад «шхщ», надзвичайно мала. Таким чином, запам'ятовуючи попередні результати, наша мережа зможе робити більш точні прогнози в майбутньому.

Досліди. Внаслідок проведених дослідів, ми отримали результат «ППРРРРР__ИИ____ВВ_І_ТТТТТ». На першій ітерації, видаляємо зайві літери: «ПР__И____В_І_Т». На 2-ій ітерації, видаляємо пропуски – « ПРИВІТ».

Висновки. Проведено дослідження технологій розпізнавання голосу на основі рекурентних нейронних мереж. Запропоновано модель розпізнавання голосу українською. Наведені результати дослідів. Отримані результати показали, що застосований алгоритм для розпізнавання голосу користувача, є ефективним. Однак, навчання мережі має такі недоліки, як потреба у великому обсягу однотипної інформації для кращого навчання мережі та точнішого розпізнавання голосу. Тобто, для досконалого розпізнавання слова «привіт» потрібно близько двадцяти голосових записів цього слова з різним тембром, швидкістю вимовляння слова та амплітудою звукової хвилі.

Список використаних джерел

1. Colletti, Justin (February 4, 2013). "The Science of Sample Rates (When Higher Is Better—And When It Isn't)". Trust Me I'm A Scientist. Retrieved February 6, 2013.

2. Bengio, Y.; Courville, A.; Vincent, P. (2013). "RepresentationLearning: AReviewandNewPerspectives". IEEETransactions on Pattern Analysis and Machine Intelligence. 1798–1828. arXiv:1206.5538. doi:10.1109/tpami.2013.50

3. Bracewell, R. N. (2000), The Fourier Transform and Its Applications (3rd ed.), Boston: McGraw-Hill, ISBN 0-07-116043-4.

4. Bengio, Yoshua; LeCun, Yann; Hinton, Geoffrey (2015). "Deep Learning". Nature. 436–444. doi:10.1038/nature14539

5. Alex Graves, Santiago Fernandez, Faustino Gomez, Jürgen Schmidhuber Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks [Electronic source] http://www.cs.toronto.edu/~graves/icml_2006.pdf

ДОВІДКА ПРО АВТОРІВ

Смішний Дмитро Миколайович – студент IV курсу, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Smishnyi Dmytro – bachelor student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: sdmitriy3758@gmail.com

Алещенко Олексій Вадимович – асистент, кафедра обчислювальної техніки, Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського".

Aleshchenko Oleksii – assistant professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: alexey.aleshchenko@gmail.com

Dmytro Smishnyi, Oleksii Aleshchenko

SPEECH RECOGNITION WITH DEEP LEARNING TECHNOLOGY USING

Task urgency. Problem of communication with a large set of devices at the same time increases too fast in our information world. It`s because each device requires its own user interface. Wherever convenient, for the user, use voice commands to manage various devices. This reduces the cost of developing UI. The instructions for such devices will be much clearer and simpler. Well, using of deep learning, in this case, cheapens and it unifies the development of voice control technology for different languages and countries. This article is devoted to the problem of voice control in Ukrainian, after all, unfortunately, there are very few such systems for this language.

Formulation of the problem. Practically complete absence of voice recognition systems for the Ukrainian segment of Internet users. The solutions that are already ready are imperfect.

Actual scientific researches and issues analysis. In recent years, you can see a significant boost in the development of voice recognition systems. The most famous examples are Amazon Alexa, Google Now, Siri, Alice from Yandex. All of them are based on neural networks. However, it is worth noting that they either do not support Ukrainian, or support for this language is extremely limited. Therefore, the article deals with the specificity of the recognition of the text in Ukrainian.

Uninvestigated parts of general matters defining. In this article the issue of recognition of voice is covered. The technology of deep learning is used. Research focuses on the generation of text based on voice, as such, and the use of this technology in relation to the Ukrainian language.

The research objective. The purpose of this article is to create a model that will work with Ukrainian words (in this case, the word Ukrainian "hello"), to recognize them, and to publish true text in accordance with the dictation.

The statement of basic materials. This article describes a method for generating text based on the Ukrainian speech. The analysis of the described stages of speech recognition based on the technology of deep learning is carried out.

Conclusions. The study of voice recognition technologies based on recurrent neural networks. A voice recognition model for Ukrainian language was proposed. The results of the experiments are presented.

Key words: recurrent neural networks, deep learning, Fourier transformation, sampling.