TECHNICAL SCIENCES AND TECHNOLOGIES
УДК 004.934:621.391:621.396.67
DOI: 10.25140/2411-5363-2017-3(9)-129-137

Олександр Корнієнко

МЕТОД ВІДОБРАЖЕННЯ МОВНИХ СИГНАЛІВ У ЗАДАЧІ РОЗПІЗНАВАННЯ МОВЦЯ

Актуальність теми дослідження. *Більшість когнітивних сервісів використовують мовні сигнали як джерело інформації, а саме: розпізнавання емоцій, мови та ідентифікація мовця. Актуальною проблемою є створення загального підходу до відображення мовних сигналів, позбавленого недоліків існуючих методів класифікації у задачі розпізнавання мовця.*

Постановка проблеми. *Більшість сучасних методів розпізнавання мовця є чутливими до тривалості мовних сигналів і, відповідно, це накладає істотні обмеження на їх застосування.*

Аналіз останніх досліджень і публікацій. *Метод зрівняння фундаментальних частот голосів та ймовірнісні підходи часто застосовують для розпізнавання мовця. Предметом більшості робіт, пов'язаних із розпізнаванням мовця, є пошук метрик зрівняння статистичних моделей голосових трактів мовців для забезпечення найвищої точності розпізнавання. Формування цих моделей (метод i-vector) здійснюється на основі статистичних розподілів короткочасних спектральних ознак. Основним недоліком такого підходу є необхідність великої кількості тренувальних даних (записів мовних сигналів великої тривалості), з метою розрахунку статистичних розподілів ознак та побудови текстонезалежної моделі мовця.*

Виділення не вирішених раніше частин загальної проблеми. *Створення загального методу виділення закономірностей у спектральних ознаках мовних сигналів короткої тривалості та характер їх зміни у часі є відкритим завданням.*

Постановка завдання. *У роботі запропоновано новий підхід до відображення мовних сигналів, як векторів ознак розподілених у часі, з використанням рекурентної нейронної мережі.*

Виклад основного матеріалу. *Розпізнавання мовця включає ідентифікацію та верифікацію людини за голосом та полягає у пошуці оптимальної пари функції відображення набору ознак мовного сигналу в багатовимірний вектор, та функції оцінки схожості таких відображень. Для пошуку альтернативної функції відображення ознак мовного сигналу в роботі використано рекурентну нейронну мережу, що складається з ланцюга двонаправлених довгих короткочасних пам'ятей. Використано евклідову відстань для спрощення процесу зрівняння зразків мовних сигналів. Для налаштування ваг рекурентної нейронної мережі використано підхід триплет втрат, що успішно використовується для розпізнавання облич.*

Висновки. *Експериментально показано, що використання запропонованого підходу дозволило зменшити помилку розпізнавання мовця EER на 7,5 % порівняно із сучасним підходом i-vector при розмірності векторів відображень 16 та 100, відповідно, для мовних сигналів тривалістю 2 с.*

Ключові слова: *розпізнавання мовця; довга короткочасна пам'ять; рекурентна нейронна мережа; підхід триплет втрат.*

Рис.: 7. Бібл.: 23.

Постановка проблеми. Задача текстонезалежного розпізнавання мовця є актуальною у сфері обробки мовних сигналів [1]. Розпізнавання особи за голосом об'єднує ідентифікацію та верифікацію мовця. Ідентифікація мовця – процес визначення особи за послідовністю ознак $\boldsymbol{x}$ мовного сигналу шляхом її порівняння з моделями голосу мовців, збереженими у базі. Результатом процесу ідентифікації є список кандидатів. Верифікація мовця полягає у перевірці запитуваної ідентичності шляхом порівняння наданої послідовності ознак $\boldsymbol{x}$ зі збереженим у базі шаблоном. Результатом верифікації є позитивне або негативне рішення.

Аналіз останніх досліджень і публікацій. Відомо кілька підходів із застосуванням нейронних мереж для пошуку оптимальних функцій $\boldsymbol{f}$ відображення та метрики зрівняння $\boldsymbol{d}$ мовних сигналів. Першим підходом до вирішення задачі ідентифікації мовця є використання багатошарового перцепторону [5] або глибинної мережі переконань (DBN, Deep Belief Network) [6]. На вхід такого алгоритму подаються ознаки мовного сигналу, наприклад, мел-частотні кепстральні коефіцієнти [7], а результатом роботи є вектор ймовірностей належності ознак до одного з класів мовців тренувальної вибірки. Однак такий метод є ресурсоємним та немасштабованим. Інший підхід полягає у пошуку функції відображення f шляхом розрахунку прихованих ознак мовного сигналу (bottle-neck features) за допомогою автоенкодеру (повнозв'язної нейронної мережі) [8]. Основне обмеження цього підходу полягає у припущенні, що глибинна нейронна мережа відобразить спектральні ознаки мовного сигналу в дикторозалежні параметри [9].

Для пошуку оптимальної функції відображення мовного сигналу як штрафну функцію у роботі використано функцію втрат триплетів мовних сигналів [10], що використ-

товується для відображення звукових записів слів в евклідовий простір [11], розділення мовців [12] та розпізнавання облич [13]. Основними відмінностями запропонованого методу є використання двонаправленої довгої короткочасної пам'яті (BLSTM, Bidirectional Long Short Term Memory) [13], функції об'єднання ознак мовного сигналу та метрики оцінки схожості послідовностей ознак мовних сигналів.

Мета статті. У роботі запропоновано новий підхід до вирішення задачі розпізнавання мовця, що базується на використанні двонаправленої рекурентної нейронної мережі для пошуку оптимальної функції відображення ***f*** та є вільним від зазначеного недоліку.

Виклад основного матеріалу. Ідентифікація та верифікація мовця є задачею мультикласової класифікації, що полягає у пошуку оптимальної пари ***(f, d)*** функцій відображення ознак мовного сигналу ***f*** у багатовимірний простір та функції оцінки схожості (метрики) ***d*** відображень зразків мовних сигналів. Для наданої послідовності ознак ***x***, відстань до відображення зразка послідовності ознак ***x+***, вимовленого цим же диктором, повинна бути меншою ніж до будь-якого іншого відображення послідовності ознак ***x-*** сигналу, вимовленого будь-яким іншим мовцем, що описується співвідношенням:

$$d(f(x), f(x_+)) < d(f(x), f(x_-)). \quad (1)$$

Метод <<i-vector>> (identity vector, вектор ідентичності) використовується як функція відображення ***f*** у сучасних системах розпізнавання мовця [2]. Об'єктом сучасних дослідження є пошук оптимальної метрики схожості ***d*** [3] для забезпечення найвищої точності розпізнавання. Основним недоліком підходу «i-vector» є чутливість до тривалості мовного сигналу [4], що накладає обмеження на формування зразків мовних сигналів та не може бути вирішений лише шляхом пошуку функції схожості.

Функція втрат триплетів мовних сигналів.

Підхід триплет втрат полягає у формуванні тренувальної вибірки триплетів послідовностей ознак (x, x_+, x_-), що відповідають представленому мовному сигналу (наданий сигнал ***x***), сигналу, вимовленому цим же мовцем (позитивний сигнал x_+) та сигналу, вимовленому будь-яким іншим мовцем (негативний сигнал x_-). Сформований триплет сигналів використовується для налаштування параметрів нейронної мережі та пошуку оптимальної функції відображення ***f***. Тренування нейронної мережі відбувається з використанням функції втрат триплетів (triplet loss function) та полягає у мінімізації відстані між відображеннями наданого та позитивного сигналів та максимізації відстані між відображеннями наданого та негативного сигналів.

Хай T – набір усіх можливих триплетів сигналів $\tau = (x_a, x_p, x_n)$ тренувальної вибірки. Функція втрат триплетів задовольняє вираз (1) та дозволяє досягти кращого розділення позитивних та негативних пар завдяки додаванню до функції втрат константи $\alpha \in \mathbb{R}^+$. Для всіх триплетів у вибірці необхідно забезпечити нерівність $\Delta_\tau + \alpha < 0$, де

$$\Delta_\tau = \|f(x_a^\tau) - f(x_a^\tau)\|_2^2 - \|f(x_a^\tau) - f(x_n^\tau)\|_2^2.$$

Налаштування параметрів нейронної мережі полягає у мінімізації функції втрат триплетів:

$$\mathcal{L}(\mathrm{T}) = \sum_{\tau \in T} \max(0, \Delta_\tau + \alpha).$$

Стратегія формування вибірок триплетів.

Як показано у роботі [10], формування всіх можливих триплетів сигналів є неефективним. Натомість для налаштування параметрів нейронної мережі використано триплети, що не задовольняють вираз $\Delta_\tau + \alpha < 0$. Усі інші триплети не вплинуть на значення функції втрат та лише збільшать обчислювальну складність алгоритму тренування. Нами використано «жорстко негативну» (hard negative) стратегію навчання [10].

Тренувальна вибірка триплетів сигналів формується для кожної епохи шляхом випадкового вибору набору ***n*** послідовностей для кожного з ***N*** мовців. Це дозволяє сформувати ***Nn(n-1)/2*** пар представлений-позитивний сигналів. Далі для кожної з цих пар випадково вибирається одна пара представлений-негативний сигнал з усіх можливих ***(N-1)n*** пар, що задовольняють нерівність $\Delta_\tau + \alpha > 0$.

Архітектура нейронної мережі.

На рис. 1 представлено структурну схему нейронної мережі, використаної для пошуку оптимальної функції відображення ***f***. Запропонована нейронна мережа складається з: ланцюга двонаправлених довгих короткочасних пам'ятей (BLSTM) розмірністю ***d1***, шару усереднення та ***L2*** нормалізації (***L2*** Average Pooling), ланцюга повнозв'язних шарів (Dense) розмірністю ***d2*** та шару ***L2*** нормалізації (L2 Normalization). Запропонована структура нейронної мережі формує одновимірний вектор розмірності (1, d2) відображення послідовності ознак мовного сигналу ***x*** розмірності ***(l, k)***, де ***l*** – кількість фреймів (тривалість послідовності), а***k*** – кількість ознак кадру мовного сигналу.

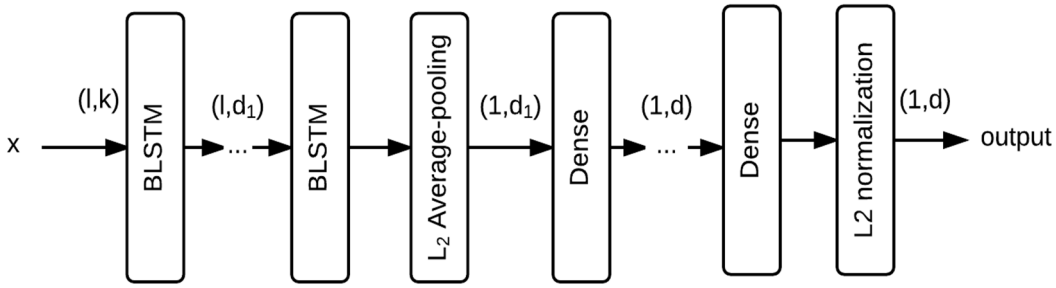

Рис. 1. Архітектура запропонованої нейронної мережі

Експериментальні результати.

Корпус мовних сигналів. Для проведення експериментів використано набір записів мовних сигналів 208 мовців загальною тривалістю близько 10 год з корпусів ASVSpoof [14] та VoxForge [15]. Кожен зразок мовного сигналу розділено на тренувальний, тестувальний сегменти та сегмент, що використовується для створення моделі мовця у співвідношенні 70, 15, 25 % відповідно від загальної тривалості запису голосу.

Ознаки мовних сигналів. Вейвлет-пакетні кепстральні коефіцієнти [16], із структурою дерева декомпозиції наближеної до психоакустичної моделі ERB [17] та базисним вейвлетом сімейства Добеші 3-го порядку використано як ознаки мовного сигналу [18]. Вектор ознак містить 18 кепстральних коефіцієнтів, їх похідних 1-го та 2-го порядку. Тривалість кадру становить 32 мс з перекриттям 16 мс. Частота дискретизації становить 16 кГц. Мовні сигнали попередньо очищенні від сегментів тиші за допомогою алгоритму [19].

Конфігурація нейронної мережі. Нейронна мережа розроблена з використанням фреймворку Keras [20] та Pyannote [12]. Евклідова метрика вибрана як міра схожості ***d***. Для пошуку оптимальної конфігурації мережі розглянуто структури з 1, 2 та 3 BLSTM, розмірністю 16, 32, 64 та 128. Розглянуто два повнозв'язні шари (Dence) з розмірністю 16, 32 та 64. Функцією активації обрано ***tanh***. Налаштування ваг нейронної мережі здійснювалось для різної тривалості мовних сигналів протягом 50 епох. Відступ ***alpha*** обрано рівним 0,2 [10]. Оптимізатором обрано модифікований алгоритм градієнтного спуску [21] зі швидкістю навчання 10^-3. Для формування триплетів використано n = 20 випадково обраних послідовностей мовних сигналів усіх мовців.

Оцінка ефективності системи розпізнавання мовця проводилась шляхом зрівняння рівня рівних помилок (EER, Error Equal Rate). Рівень рівних помилок представляє величину ймовірності помилок при такому порозі, при якому ймовірність помилок 1-го та 2-го роду збігаються або близькі за значенням.

Як альтернативний метод розпізнавання мовця обрано підхід «i-vector». Модель мовця створювалась з використанням програмного пакета BOB [22] з такими параметрами: розмірність «вектора ідентичності» (identity vector, i-vector) 100, кількість компонент суміші Гауссових розподілів 256, PLDA класифікатор із розмірністю векторів лінійних моделей 50.

На рис. 2 представлено залежність помилки розпізнавання мовця EER від епохи тренування для тестової вибірки мовних сигналів тривалістю 2 с. Як бачимо, точність розпізнавання мовця збільшується при збільшенні кількості епох тренування нейронної мережі.

Рис. 2. Залежність помилки розпізнавання мовця EER від епохи тренування (тривалість мовних сигналів 2 с, кількість LSTM 1, d1=32, d2=16)

На рис. 3 зображені t-SNE [23] проекції векторів відображень сигналів 35 мовців тестової вибірки. Таким чином, більшість векторів відображень чітко розділені на групи та формують класи мовців.

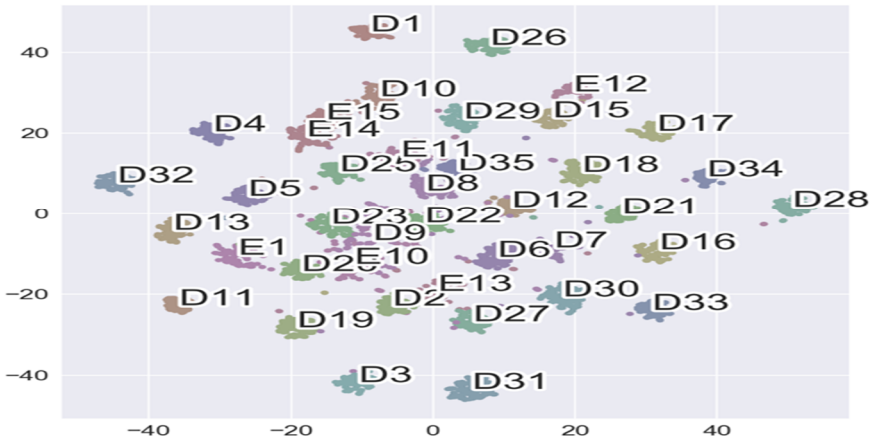

Рис. 3. t-SNE проекція відображень сигналів 35 мовців

Точність запропонованого методу розпізнавання мовця є вищою на 7,6 % ніж для методу «i-vector», що представлено залежністю EER від тривалості мовного сигналу (рис. 4).

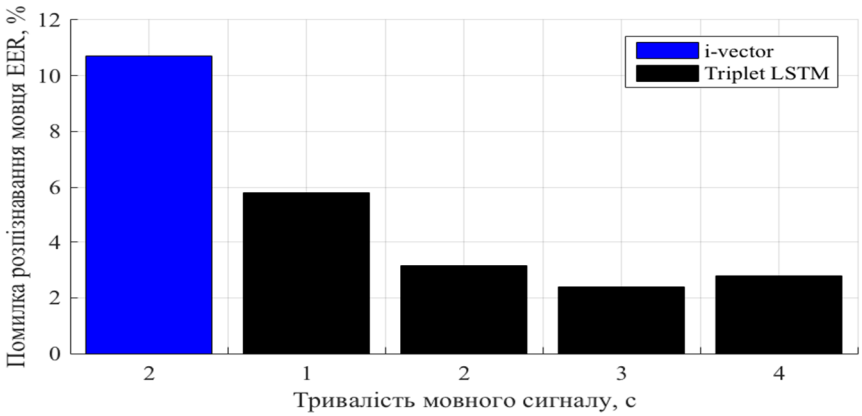

Рис. 4. Залежність помилки EER розпізнавання мовця від тривалості мовного сигналу запропонованою системою Triplet LSTM (d1=32, d=16, nlstm=1, ndense=2) та «i-vector»

Залежності помилки розпізнавання EER від кількості та розмірності BLSTM представлено на рис. 5 та 6 відповідно. Найбільша точність розпізнавання мовця досягається для d1 = 32 та Nblstm = 1. При збільшенні кількості ланцюгів рекурентної нейронної мережі спостерігається перенавчання, тобто нейронна мережа «запам'ятовує» змістову частину повідомлення, що зумовлює падіння точності текстонезалежного розпізнавання мовця.

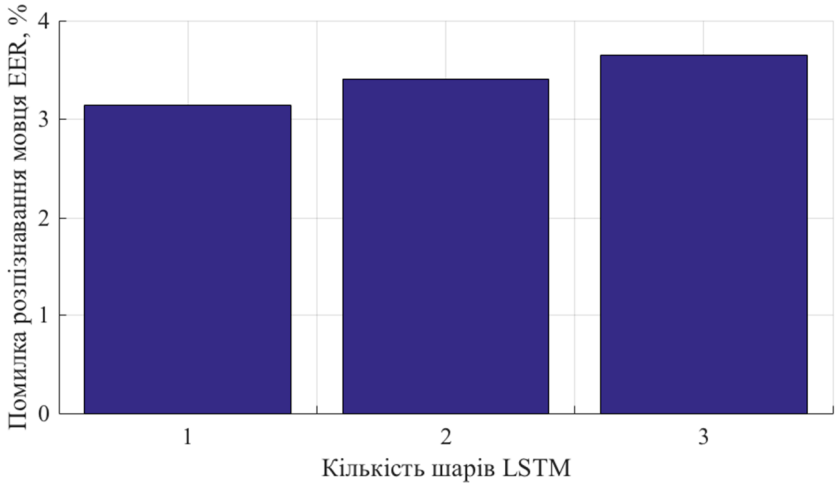

Рис. 5. Залежність помилки EER розпізнавання мовця від кількості шарів LSTM запропонованою системою (t=2c, d1=32, d=16, ndense=2)

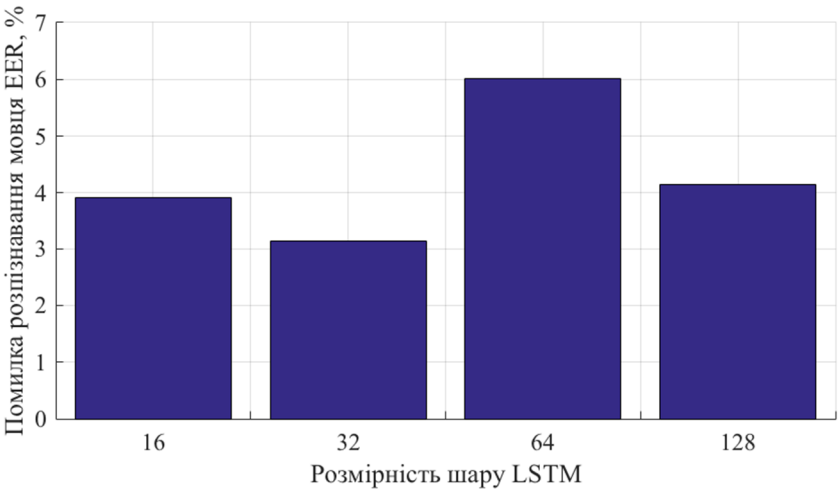

Рис. 6. Залежність помилки EER розпізнавання мовця від розмірності шару LSTM запропонованою системою (t=2c, nlstm=1, d=16, ndense=2)

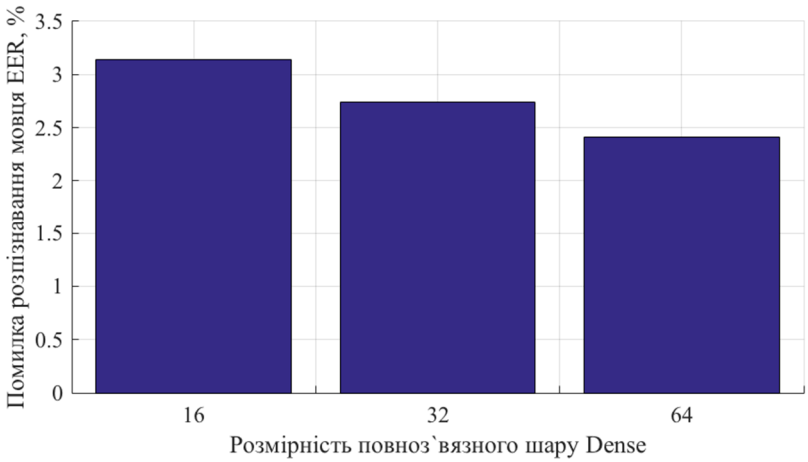

Рис. 7. Залежність помилки EER розпізнавання мовця від розмірності шару Dense запропонованою системою (t=2c, nlstm=1, d1=32, ndense=2)

Отже, експериментально визначено параметри запропонованої архітектури нейронної мережі, для яких досягається найбільша ймовірність правильної класифікації мовців 8,4 %:

- Кількість ланцюгів рекурентної нейронної мережі BLSTM: 1.
- Розмірність ланцюга рекурентної нейронної мережі: 32.
- Розмірність повнозв'язного шару Dense: 64.

Висновки та пропозиції. Запропоновано метод пошуку оптимальної функції відображення послідовності ознак мовного сигналу в завданні розпізнавання мовців. Визначено, що помилка розпізнавання мовця EER запропонованого методу на 7,5 % менша порівняно з актуальним підходом «i-vector» при розмірності векторів відображень мовних сигналів 16 та 100 відповідно. Визначено оптимальні параметри архітектури нейронної мережі: розмірність ланцюга BLSTMd1 = 32, кількість ланцюгів BLSTM Nblstm = 1, розмірність повнозв'язного шару Dence d = 16. Виявлено, що ускладнення моделі нейронної мережі шляхом збільшення шарів LSTM приводить до перенавчання (запам'ятовування змістової складової повідомлення) та, відповідно, зменшення точності текстонезалежного розпізнавання мовця. Становить інтерес оцінки точності розпізнавання запропонованого методу на корпусах більшої ємності, наприклад LibriSpeech, а також використання інших типів метрики схожості та структур нейронних мереж. Запропонований метод може бути використаний для вирішення завдань текстонезалежного розпізнавання мовця, розділення мовців (speaker diarization), розпізнавання емоцій та щодо інших типів класифікації мовних сигналів.

Список використаних джерел

1. Kinnunen, Tomi, Li Haizhou (2010). An Overview of Text-independent Speaker Recognition: From Features toSupervectors. *Speech Commun*, vol. 52, no. 1, pp. 12–40. Retrieved from http://dx.doi.org/10.1016/j.specom.2009.08.009.

2. Dehak, N., Kenny, P. J. & Dehak R. et al. (2011). Front-End Factor Analysis for Speaker Verification. *IEEETransactions on Audio, Speech, and Language Processing,* vol. 19, no. 4, pp. 788–798.

3. Garcia-Romero, Daniel, Espy-Wilson Carol Y. (2011). Analysis of i-vector Length Normalization in Speaker Recognition Systems. *Interspeech*, vol. 2011, pp. 249–252.

4. Achintya, Kumar Sarkar, Driss, Matrouf (eds.) (2012). Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification. *INTERSPEECH. ISCA*, pp. 2662–2665. Retrieved from http://dblp.uni-trier.de/db/conf/interspeech/interspeech2012.html#SarkarMBB12.

5. Ehsan, Variani, Xin, Lei & Erik McDermott et al. (2014). Deep neural networks for small footprint text-dependent speaker verification. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE InternationalConference on / IEEE*, pp. 4052–4056.

6. Ghahabi, Omid, Hernando, Javier (2014). Deep belief networks for i-vector based speaker recognition. *Acoustics,Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on / IEEE*, pp.1700–1704.

7. Davis Steven, Mermelstein Paul (1980). Comparison of parametric representations for monosyllabic word recognitionin continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366.

8. Richardson, Fred, Reynolds, Douglas & Dehak Najim (2015). Deep neural network approaches to speaker and languagerecognition . *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675.

9. Yella Sree Harsha, Stolcke Andreas, Slaney Malcolm (2014). Artificial neural network features for speaker diarization. *Spoken Language Technology Workshop (SLT), 2014 IEEE / IEEE*, pp. 402–406.

10. Schroff Florian, Kalenichenko Dmitry, Philbin James (2015). Facenet: A unified embedding for face recognition andclustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823.

11. He Wanjia, Wang Weiran, Livescu Karen (2016). Multi-view Recurrent Neural Acoustic Word Embeddings. *CoRR*, vol. abs/1611.04496. Retrieved from http://arxiv.org/abs/1611.04496.

12. Bredin, Hervé (2016) TristouNet: Triplet Loss for Speaker Turn Embedding. *CoRR*, vol. abs/1609.04301. Retrieved from http://arxiv.org/abs/1609.04301.

13. Sundermeyer Martin, Schlüter Ralf & Ney Hermann. (2012). LSTM Neural Networks for Language Modeling. *Interspeech*, pp. 194–197.

14. Zhizheng Wu, Tomi Kinnunen & Nicholas Evans et al. (2015). ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. *Training*, vol. 10, no. 15, pp. 3750.

15. *VoxForge project.* Retrieved from http://voxforge.org.

16. Sarikaya Ruhi, Pellom Bryan L., & Hansen John HL. (1998). Wavelet packet transform features with application to speaker identification. *IEEE Nordic signal processing symposium / CiteSeerX*, pp. 81–84.

17. Moore, Brian CJ. (2012). *An introduction to the psychology of hearing*. Brill.

18. Корнієнко О. О. Вейвлет-пакетні ознаки мовного сигналу у завданні розпізнавання мовця / О. О. Корнієнко // Вимірювальна та обчислювальна техніка в технологічних процесах : міжнар. наук.-техн. журн. – 2017. – № 2. – С. 111–117.

19. Alam, J. Patrick Kenny & Pierre Ouellet et al. (2014). Supervised/Unsupervised Voice Activity Detectors for Textdependent Speaker Recognition on the RSR2015 Corpus. *Odyssey Speaker and Language Recognition Workshop*. Retrieved from http://www.crim.ca/perso/patrick.kenny/Alam_odyssey2014.pdf.

20. Chollet François. Keras. 2015.

21. Funk, Simon (2015). RMSprop loses to SMORMS3 – beware the epsilon! Retrieved from http://sifter.org/~simon/journal/20150420.html.

22. Khoury Elie, El Shafey Laurent & Marcel Sébastien (2014). Spear: An open source toolbox for speaker recognition based on Bob. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on / IEEE*, pp. 1655–1659.

23. Maaten Laurens van der, Hinton Geoffrey. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, vol. 9, no. Nov., pp. 2579–2605.

References

1. Kinnunen, Tomi, Li Haizhou (2010). An Overview of Text-independent Speaker Recognition: From Features toSupervectors. *Speech Commun*, vol. 52, no. 1, pp. 12–40. Retrieved from http://dx.doi.org/10.1016/j.specom.2009.08.009.

2. Dehak, N., Kenny, P. J. & Dehak R. et al. (2011). Front-End Factor Analysis for Speaker Verification. *IEEETransactions on Audio, Speech, and Language Processing,* vol. 19, no. 4, pp. 788–798.

3. Garcia-Romero, Daniel, Espy-Wilson Carol Y. (2011). Analysis of i-vector Length Normalization in Speaker Recognition Systems. *Interspeech*, vol. 2011, pp. 249–252.

4. Achintya, Kumar Sarkar, Driss, Matrouf (eds.) (2012). Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification. *INTERSPEECH.*

ISCA, pp. 2662–2665. Retrieved from http://dblp.uni-trier.de/db/conf/interspeech/interspeech2012.html#SarkarMBB12.

5. Ehsan, Variani, Xin, Lei & Erik McDermott et al. (2014). Deep neural networks for small footprint text-dependent speaker verification. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE InternationalConference on / IEEE*, pp. 4052–4056.

6. Ghahabi, Omid, Hernando, Javier (2014). Deep belief networks for i-vector based speaker recognition. *Acoustics,Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on / IEEE*, pp. 1700–1704.

7. Davis Steven, Mermelstein Paul (1980). Comparison of parametric representations for monosyllabic word recognitionin continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366.

8. Richardson, Fred, Reynolds, Douglas & Dehak Najim (2015). Deep neural network approaches to speaker and languagerecognition . *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675.

9. Yella Sree Harsha, Stolcke Andreas, Slaney Malcolm (2014). Artificial neural network features for speaker diarization . *Spoken Language Technology Workshop (SLT), 2014 IEEE / IEEE*, pp. 402–406.

10. Schroff Florian, Kalenichenko Dmitry, Philbin James (2015). Facenet: A unified embedding for face recognition andclustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823.

11. He Wanjia, Wang Weiran, Livescu Karen (2016). Multi-view Recurrent Neural Acoustic Word Embeddings. *CoRR*, vol. abs/1611.04496. Retrieved from http://arxiv.org/abs/1611.04496.

12. Bredin, Hervé. (2016) TristouNet: Triplet Loss for Speaker Turn Embedding. *CoRR*, vol. abs/1609.04301. Retrieved from http://arxiv.org/abs/1609.04301.

13. Sundermeyer Martin, Schlüter Ralf & Ney Hermann (2012). LSTM Neural Networks for Language Modeling. *Interspeech*, pp. 194–197.

14. Zhizheng Wu, Tomi Kinnunen & Nicholas Evans et al. (2015). ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. *Training*, vol. 10, no. 15, pp. 3750.

15. VoxForge project. Retrieved from http://voxforge.org.

16. Sarikaya Ruhi, Pellom Bryan L., & Hansen John HL. (1998). Wavelet packet transform features with application to speaker identification. *IEEE Nordic signal processing symposium / CiteSeerX*, pp. 81–84.

17. Moore, Brian CJ. (2012). *An introduction to the psychology of hearing*. Brill.

18. Korniyenko O. O. (2017) Veyvlet-paketni oznaky movnoho syhnalu u zavdanni rozpiznavannya movtsya [Wavelet and package features of voice signal concerning recognition speaker problem]. Vymiriuvalna ta obchysliuvalna tekhnika v tekhnolohichnykh protsesakh – Measuring and computing engineering in technological processes, no 2, pp. 98–104 (in Ukrainain).

19. Alam, J. Patrick Kenny & Pierre Ouellet et al. (2014). Supervised/Unsupervised Voice Activity Detectors for Textdependent Speaker Recognition on the RSR2015 Corpus. *Odyssey Speaker and Language Recognition Workshop*. Retrieved from http://www.crim.ca/perso/patrick.kenny/Alam_odyssey2014.pdf.

20. Chollet François. Keras. 2015.

21. Funk, Simon (2015). RMSprop loses to SMORMS3 - beware the epsilon! Retrieved from http://sifter.org/~simon/journal/20150420.html.

22. Khoury Elie, El Shafey Laurent & Marcel Sébastien (2014). Spear: An open source toolbox for speaker recognition based on Bob. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on / IEEE*, pp. 1655–1659.

23. Maaten Laurens van der, Hinton Geoffrey. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, vol. 9, no. Nov., pp. 2579–2605.

TECHNICAL SCIENCES AND TECHNOLOGIES

UDC 004.934:621.391:621.396.67

Oleksandr Korniienko

THE METHOD OF DISPLAYING SPEECH SIGNALS IN THE SPEECH RECOGNITION TASK

Urgency of the research. *Most cognitive services deal with voice understanding of emotions, speech and speaker recognition. Thus, the actual problem is creating of general approach for speech embedding, such as speaker recognition.*

Target setting. *The state-of-art speaker recognition methods have significant restrictions on their use because these methods are sensitive to durations of the speech signals.*

Actual scientific researches and issues analysis. *The method of comparison fundamental frequency and probabilistic approaches are often used for speaker recognition. The subject of the majority of speaker recognition researches is searching the metrics for similarity scoring of voice statistical models. The main goal of these researches is to ensure the highest accuracy of the recognition. Researches propose to generate voice models using the probability distributions of short-term spectral features. This method is called i-vector**.*** *The main disadvantages of statistical models are that they must have large training speech corpus to calculate the statistical distributions of features and construct a text-independent model of speaker.*

Uninvestigated parts of general matters defining. *Creating a general method for patterns extraction from time-distributed short-term spectral features is required.*

The research objective*. In this paper we proposed a new approach to the speech signals embedding using a recurrent neural network, which can be used for speaker, speech and emotion recognition.*

The statement of basic materials. *Speaker recognition involves the identification of a person and verification by the voice and boils down to find the optimal pairs of speech signal representing function, and scoring function for evaluating the similarity between given and known speech signals. In order to find an alternative function of speech signal embedding, a bidirectional long short-term memory is used. The euclidean distance is used to simplify the process of measuring the similarity between speech signals. The triplet loss function is minimized for adjust the weights of the recurrent neural network. This is because the optimization approach is successfully used for face recognition.*

Conclusions. *It has been shown experimentally that the use of the proposed approach allowed to reduce the speaker recognition error equal rate by 7.5 % compared with the state-of-art i-vector approach with voice models vector dimension 16 and 100, respectively, for 2 sec. speech signals.*

Key words: *speaker recognition; long short-term memory; reccurent neural network; triplet loss function.*

Fig.: 7. Bibl.: 23.

УДК 004.934:621.391:621.396.67

Александр Корниенко

МЕТОД ОТОБРАЖЕНИЯ РЕЧЕВЫХ СИГНАЛОВ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ГОВОРЯЩЕГО

Большинство когнитивных сервисов используют речевые сигналы как источник информации, а именно: распознавание эмоций, речи и идентификация говорящего. Актуальной проблемой является создание общего подхода к отражению речевых сигналов, лишенного недостатков существующих методов классификации в задаче распознавания говорящего. В работе предложен новый подход к отражению языковых сигналов, как векторов признаков распределенных во времени, с использованием рекуррентной нейронной сети.

Для поиска альтернативной функции отображения признаков речевого сигнала в работе использовано рекуррентную нейронную сеть, состоящую из цепи двунаправленных долгих кратковременных памятей. Использовано эвклидово расстояние для упрощения процесса уравнивания образцов речевых сигналов. Для настройки весов рекуррентной нейронной сети использован подход триплет потерь, что успешно используется для распознавания лиц.

Экспериментально показано, что использование предложенного подхода позволило уменьшить ошибку распознавания говорящего EER на 7,5 % по сравнению с современным подходом i-vector при размерности векторов отображений 16 и 100, соответственно, для речевых сигналов длительностью 2 с.

Ключевые слова: *распознавание говорящего; длинная кратковременная память; рекуррентная нейронная сеть; подход триплет потерь.*

Рис.: 7. Библ.: 23.

Корнієнко Олександр Олегович – аспірант, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» (просп. Перемоги, 37, 03056, м. Київ, Україна).

Корниенко Александр Олегович – аспирант, Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского» (просп. Победы, 37, 03056, г. Киев, Украина).

Korniienko Oleksandr – PhD student, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" (37 Pobedy Av., 03056 Kyiv, Ukraine).

E-mail: olexandr.korniienko@gmail.com

Корнієнко О. Метод відображення мовних сигналів у задачі розпізнавання мовця / О. Корнієнко // Технічні науки та технології. – 2017. – № 3 (9). – С. 129-137.