

МЕТОД ВІДОБРАЖЕННЯ МОВНИХ СИГНАЛІВ У ЗАДАЧІ РОЗПІЗНАВАННЯ МОВЦЯ

О. О. Корнієнко^{1, а}

¹ *Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»,
Фізико-технічний інститут*

Анотація

Розроблено алгоритм відображення послідовності ознак мовного сигналу в евклідовий простір. Показано, що використання рекурентної нейронної мережі для пошуку функції відображення дозволило підвищити точність розпізнавання мовця на 2% порівнянно з актуальним методом «i-vector».

Ключові слова: розпізнавання мовця, відображення мовних сигналів, LSTM

Вступ

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 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 x із збереженим у базі шаблоном. Результатом верифікації є позитивне або негативне рішення.

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.

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.

1. Пов'язані роботи

Відомо кілька підходів із застосуванням нейронних мереж для пошуку оптимальних функцій f відображення та метрики зрівняння d мовних сигналів. Першим підходом до вирішення задачі ідентифікації мовця є використання багатопарового 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 приналежності ознак до одного з класів мовців тренувальної вибірки. Однак такий метод є ресурсоемним та немасштабованим. Інший підхід полягає у пошуку функції відображення f шляхом розрахунку прихованих ознак мовного сигналу (bottle-neck features) [8] за допомогою повнозв'язної нейронної мережі. Основне обмеження цього підходу полягає у припущенні, що глибинна нейронна мережа відобразить спектральні ознаки мовного сигналу у дикторозалежні параметри [9].

Для пошуку оптимальної функції відображення мовного сигналу, як штрафну функцію у роботі використано функцію втрат триплетів мовних сигналів [10], що використовується для відображення звукових записів слів в евклідовий простір [11], розділення мовців [12] та розпізнавання обличчя [10]. 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.

^аolexandr.korniienko@gmail.com

2. 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 $\tau = (x, x_+, x_-)$ тренувальної вибірки. Функція втрат триплетів задовольняє вираз (1) та дозволяє досягти кращого розділення позитивних та негативних пар завдяки додаванню до функції втрат константи $\alpha \in R^+$. Для всіх триплетів у вибірці необхідно забезпечити нерівність $\Delta_\tau + \alpha < 0$, де

$$*\Delta_\tau = d(f(x), f(x_+)) - d(f(x), f(x_-)) \quad (2)$$

Налаштування параметрів нейронної мережі полягає у мінімізації функції втрат триплетів:

$$*L(T) = \sum_{\tau \in T} \max(0, \Delta_\tau + \alpha) \quad (3)$$

2.1. 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 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 $\Delta_\tau + \alpha > 0$.

3. 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 d_1 , шару усереднення та L_2 нормалізації (L_2 Average

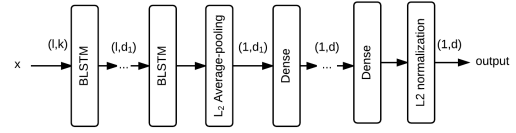


Рис. 1. Структурна схема нейронної мережі

Pooling), ланцюга повнозв'язних шарів (Dense) розмірністю d_2 та шару L_2 нормалізації (L_2 Normalization). Запропонована структура нейронної мережі формує одновимірний вектор розмірності $(1, d_2)$ відображення послідовності ознак мовного сигналу x розмірності (l, k) , де l – тривалість послідовності, k – кількість ознак кадру мовного сигналу.

4. Експериментальні результати

Корпус мовних сигналів. Для проведення експериментів використано набір записів мовних сигналів 40 мовців загальною тривалістю 5 год з корпусу ASVSpoof [14]. Набір розділено на тренувальну, валідаційну та тестову вибірки у співвідношенні (70%, 10%, 20%) від загальної тривалості. Загалом у тренувальній вибірці були записи мовних сигналів 30 мовців, валідаційній – 35 мовців (з яких 5 нових) та тестовій вибірці – 40 мовців (з яких 10 нових).

Ознаки мовних сигналів. Вейвлет-пакетні кепстральні коефіцієнти [15], із структурою дерева декомпозиції наближеної до психоакустичної моделі ERB [16] та базисним вейвлетом сімейства Добеші 3-ого порядку використано як ознаки мовного сигналу. Вектор ознак містить 18 кепстральних коефіцієнтів, їх похідних 1-ого та 2-ого 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 [17].

Конфігурація нейронної мережі. Нейронна мережа розроблена з використанням фреймворку Keras [18]. Евклідова метрика вибрана як міра схожості d . Для пошуку оптимальної конфігурації мережі розглянуто структури з 1 та 2 BLSTM, розмірністю 8, 16 та 32. Розглянуто два повнозв'язні шари (Dense) розмірністю 16. Функцією активації обрано \tanh . Кожна модель тренувалась для різної тривалості мовних сигналів впродовж 50 епох. Відступ α обрано рівним 0.2 [10]. Оптимізатором обрано модифікований алгоритм градієнтного спуску [19] зі швидкістю навчання 10^{-3} . Для формування триплетів використано $n = 20$ випадково обраних послідовностей для кожного мовця.

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-ого та 2-ого роду співпадають або близькі по значенню.

Як альтернативний метод розпізнавання мовця обрано підхід «i-vector». Модель мовця створювалась з використанням програмного пакету BOV [20] з такими параметрами: розмірність “вектору ідентичності” (identity vector, i-vector) 100, кількість ком-

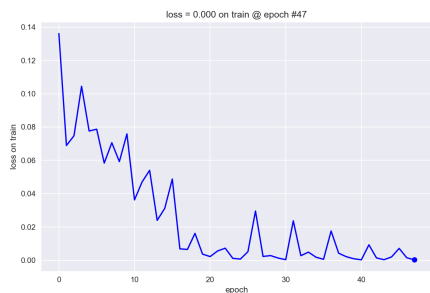


Рис. 2. Залежність значення функції втрат $L(T)$ від епохи тренування

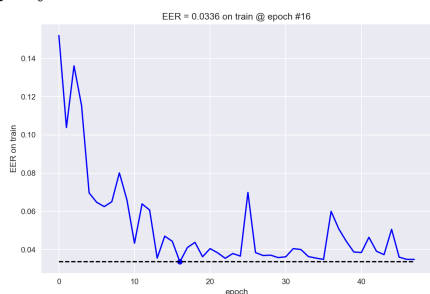


Рис. 3. Залежність помилки EER від епохи тренування



Рис. 4. t-SNE проекції відображень сигналів мовців понент суміші Гаусових розподілів 256, PLDA класифікатор із розмірністю векторів лінійних моделей 50.

На рис. 2, 3 представлено залежності значення функції втрат триплетів мовних сигналів тривалістю 2 с та помилки EER від епохи тренування для валідаційної вибірки. Як бачимо, точність розпізнавання мовця збільшується на кожному кроці тренування нейронної мережі.

На рис. 4 зображені t-SNE [21] проекції векторів відображень сигналів мовців тестової вибірки. Таким чином, більшість векторів відображень чітко розділені на групи та формують класи мовців.

Точність запропонованого методу розпізнавання мовця є вищою на 2 % ніж для методу «i-vector», що представлено залежністю EER від тривалості мовного сигналу (рис. 5).

Залежності помилки розпізнавання EER від розмірності та кількості BLSTM представлено на

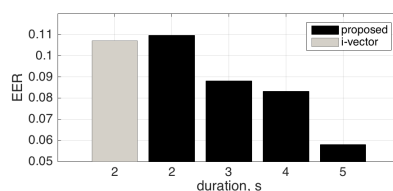


Рис. 5. Залежність EER від тривалості мовного сигналу. $d_1 = 32, N_{blstm} = 1$

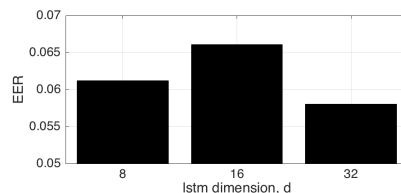


Рис. 6. Залежність EER від розмірності BLSTM. $N_{blstm} = 1$

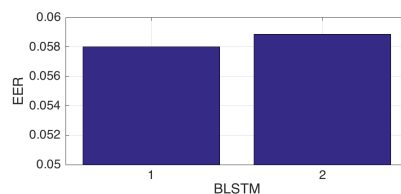


Рис. 7. Залежність EER від кількості BLSTM. $d_1 = 32$

рис. 6 та 7 відповідно. Найбільша точність розпізнавання мовця досягається для $d_1 = 32$ та $N_{blstm} = 1$. При збільшенні складності моделі, спостерігається перенавчання, що зумовлює падіння точності розпізнавання мовця.

Висновки

Запропоновано метод пошуку оптимальної функції відображення послідовності ознак мовного сигналу у завданні розпізнавання мовців. Визначено, що помилка розпізнавання EER запропонованого методу на 2% менша порівняно з актуальним підходом «i-vector» при розмірності векторів відображень мовних сигналів 16 та 100 відповідно. Визначено оптимальні параметри нейронної мережі: $d_1 = 32, N_{blstm} = 1$. Виявлено, ускладнення моделі нейронної мережі приводить до зменшення точності розпізнавання мовця. Представляє інтерес оцінки точності розпізнавання запропонованого методу на корпусах більшої ємності, наприклад NIST, а також використання інших типів метрики схожості та структури нейронної мережі. Запропонований метод може бути використаний для вирішення завдань текстонезалежного розпізнавання мовця, розділення мовців (speaker diarization), розпізнавання емоцій та щодо інших типів класифікації мовних сигналів.

Перелік використаних джерел

1. Kinnunen Tomi, Li Haizhou. An Overview of Text-independent Speaker Recognition: From Features to Supervectors // Speech Commun. — 2010. — Jan. — Vol. 52, no. 1. — P. 12–40. — Access

- mode: <http://dx.doi.org/10.1016/j.specom.2009.08.009>.
2. Front-End Factor Analysis for Speaker Verification / N. Dehak, P. J. Kenny, R. Dehak et al. // IEEE Transactions on Audio, Speech, and Language Processing. — 2011. — May. — Vol. 19, no. 4. — P. 788–798.
 3. Garcia-Romero Daniel, Espy-Wilson Carol Y. Analysis of i-vector Length Normalization in Speaker Recognition Systems. // Interspeech. — Vol. 2011. — 2011. — P. 249–252.
 4. Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification. / Achintya Kumar Sarkar, Driss Matrouf, Pierre-Michel Bousquet, Jean-François Bonastre // INTERSPEECH. — ISCA, 2012. — P. 2662–2665. — Access mode: <http://dblp.uni-trier.de/db/conf/interspeech/interspeech2012.html#SarkarMBB12>.
 5. Deep neural networks for small footprint text-dependent speaker verification / Ehsan Variiani, Xin Lei, Erik McDermott et al. // Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on / IEEE. — 2014. — P. 4052–4056.
 6. Ghahabi Omid, Hernando Javier. Deep belief networks for i-vector based speaker recognition // Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on / IEEE. — 2014. — P. 1700–1704.
 7. Davis Steven, Mermelstein Paul. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences // IEEE transactions on acoustics, speech, and signal processing. — 1980. — Vol. 28, no. 4. — P. 357–366.
 8. Richardson Fred, Reynolds Douglas, Dehak Najim. Deep neural network approaches to speaker and language recognition // IEEE Signal Processing Letters. — 2015. — Vol. 22, no. 10. — P. 1671–1675.
 9. Yella Sree Harsha, Stolcke Andreas, Slaney Malcolm. Artificial neural network features for speaker diarization // Spoken Language Technology Workshop (SLT), 2014 IEEE / IEEE. — 2014. — P. 402–406.
 10. Schroff Florian, Kalenichenko Dmitry, Philbin James. Facenet: A unified embedding for face recognition and clustering // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2015. — P. 815–823.
 11. He Wanjia, Wang Weiran, Livescu Karen. Multi-view Recurrent Neural Acoustic Word Embeddings // CoRR. — 2016. — Vol. abs/1611.04496. — Access mode: <http://arxiv.org/abs/1611.04496>.
 12. Bredin Hervé. TristouNet: Triplet Loss for Speaker Turn Embedding // CoRR. — 2016. — Vol. abs/1609.04301. — Access mode: <http://arxiv.org/abs/1609.04301>.
 13. Sundermeyer Martin, Schlüter Ralf, Ney Hermann. LSTM Neural Networks for Language Modeling. // Interspeech. — 2012. — P. 194–197.
 14. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge / Zhizheng Wu, Tomi Kinnunen, Nicholas Evans et al. // Training. — 2015. — Vol. 10, no. 15. — P. 3750.
 15. Sarikaya Ruhi, Pellom Bryan L, Hansen John HL. Wavelet packet transform features with application to speaker identification // IEEE Nordic signal processing symposium / CiteSeerX. — 1998. — P. 81–84.
 16. Moore Brian CJ. An introduction to the psychology of hearing. — Brill, 2012.
 17. Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the rsr2015 corpus / J Alam, Patrick Kenny, Pierre Ouellet et al. // Odyssey Speaker and Language Recognition Workshop. — 2014.
 18. Chollet François. Keras. — 2015.
 19. Funk Simon. RMSprop loses to SMORMS3 - beware the epsilon! — 2015. — Access mode: <http://sifter.org/~simon/journal/20150420.html>.
 20. Khoury Elie, El Shafey Laurent, Marcel Sébastien. Spear: An open source toolbox for speaker recognition based on Bob // Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on / IEEE. — 2014. — P. 1655–1659.
 21. Maaten Laurens van der, Hinton Geoffrey. Visualizing data using t-SNE // Journal of Machine Learning Research. — 2008. — Vol. 9, no. Nov. — P. 2579–2605.