УДК 004.852:004.93’1

М.В. Панченко[1], П.І. Сагайда[1]
[1]ТОВ «ТЕХНІЧНИЙ УНІВЕРСИТЕТ «МЕТІНВЕСТ ПОЛІТЕХНІКА», Запоріжжя, Україна

ECAPA-TDNN ДЛЯ ГОЛОСОВОЇ БІОМЕТРІЇ: ПЕРЕВАГИ НАД X-VECTORS ТА RESNET ДЛЯ ІНТЕГРОВАНОЇ ВЕРИФІКАЦІЇ МОВЦЯ ТА АНТИСПУФІНГУ

Анотація. У роботі представлено концепт інтегрованої системи голосової біометрії, що поєднує верифікацію мовця (SV) та захист від спуфінгових атак (AS) на основі єдиної архітектури ознак ECAPA-TDNN. Запропонований уніфікований підхід дозволяє створювати легші, більш стійкі до шумів та ефективніші системи, що критично важливо для протидії сучасним Deepfake-атакам та спрощує масштабованість біометричної платформи.

Ключові слова: *голосова біометрія, верифікація мовця, антиспуфінг, ECAPA-TDNN, нейронні мережі, ембеддинги мовлення*

Вступ. Верифікація мовця (Speaker Verification, SV) та захист від спуфінгу (Anti-Spoofing, AS) є ключовими складовими сучасних голосових біометричних систем, оскільки голос дедалі частіше використовується як зручний і природний засіб автентифікації. SV забезпечує перевірку заявленої особи шляхом зіставлення поточного голосового сигналу з еталонним зразком, що дозволяє реалізувати безпечний доступ до сервісів і персоналізованих функцій у банківських, мобільних та IoT-системах. Натомість AS спрямований на виявлення та блокування спуфінгових атак, коли зловмисники намагаються імітувати голос користувача за допомогою відтворення записів, синтезованого мовлення або технологій перетворення голосу.

Постановка задачі. Дослідження та порівняння ефективності моделей ECAPA-TDNN та класичних підходів, зокрема x-vectors і ResNet, для інтегрованої системи голосової біометрії, яка виконує одночасно верифікацію мовця та захист від спуфінгу. Основним підходом є всебічний аналіз можливостей кожної архітектури з метою виявлення її здатності забезпечувати високу точність, робастність до шумів, акустичних варіацій та сучасних атак Deepfake.

Основний зміст роботи. TDNN-базований екстрактор x-векторів забезпечує отримання фіксованих векторних представлень аудіосегментів довільної тривалості та вирізняється високою ефективністю у задачах верифікації мовця. Його архітектура побудована на time-delay neural network, яка агрегує локальні часові залежності у мовному сигналі та дає змогу опрацьовувати мовні фрагменти різної довжини без втрати структурної інформації. Завдяки цьому x-вектори стали практичним і широко застосовуваним базовим підходом у багатьох системах SV [1].

У сучасних дослідженнях наголошується, що для забезпечення високої робастності у реальних акустичних умовах архітектурі TDNN необхідні

додаткові механізми контекстного моделювання. Зокрема, автори підкреслюють, що традиційна структура TDNN, яка лежить в основі x-векторів, має обмеження в здатності охоплювати довготривалий контекст мовлення. [2]

ResNet-базовані моделі формують глибші та більш виразні ембеддинги мовного сигналу завдяки використанню розгалуженої архітектури згорткових нейронних мереж та залишкових блоків. Завдяки їм мережа здатна навчати багаторівневі ознаки, долаючи проблему затухання градієнтів у глибоких архітектурах і забезпечуючи більш стабільне тренування [3]. У контексті завдань SV та AS такий підхід дозволяє ефективніше моделювати складні акустичні закономірності та підвищувати дискримінативність ембеддингів.

Однак збільшення глибини та кількості параметрів робить ResNet-моделі більш схильними до перенавчання, особливо у випадках, коли тренувальні дані містять значний рівень шумів або мають нерівномірну якість.

ECAPA-TDNN є удосконаленою архітектурою на основі TDNN, яка поєднує механізми каналової уваги Squeeze-and-Excitation (SE), багатомасштабні Res2Net-блоки та покращені методи агрегації ознак. Завдяки SE-модулям модель навчається виділяти найбільш інформативні канали, тоді як Res2Net забезпечує багаторівневе представлення часових залежностей, що дозволяє враховувати різні масштаби акустичної інформації. Крім того, вдосконалена система статистичної агрегації з увагою по каналах підсилює значущі фрейми для кожної ознаки, що робить отримані ембеддинги суттєво більш дискримінативними та стійкими у завданнях верифікації мовця.

Такі архітектурні вдосконалення забезпечують істотний приріст точності порівняно з традиційними TDNN та CNN-підходами, особливо у сценаріях із наявністю шумів, спотворень та міжмовних варіацій. ECAPA-TDNN стала сучасним стандартом для побудови високоточних систем SV/AS, поєднуючи високу робастність та здатність до узагальнення [4].

Наукова новизна. Запропонований концепт інтегрованої системи верифікації мовця та антиспуфінгу передбачає використання ECAPA-TDNN як єдиної базової архітектури для екстракції голосових ознак. Хоча модель вже зарекомендувала себе як провідна архітектура для окремої задачі верифікації мовця [4], на відміну від попередніх інтегрованих систем на базі x-vector, архітектура ECAPA-TDNN завдяки механізмам уваги та Res2Net забезпечує формування більш узгодженого та дискримінативного простору ембеддингів для обох задач. Це, як очікується, дозволить істотно підвищити стійкість до акустичних варіацій та сучасних атак Deepfake [5]. Крім того, збереження легкості та мінімізація ризику перенавчання є ключовою перевагою над глибокими CNN-архітектурами типу ResNet. На основі цих уніфікованих представлень пропонується побудова двох паралельних модулів (SV та AS), що підвищує масштабованість, надійність та ефективність тренування компонентів системи на різних наборах даних (наприклад, VoxCeleb для SV та ASVspoof для AS).

Висновки. ECAPA-TDNN забезпечує формування більш дискримінативних і робастних голосових ембеддингів завдяки механізмам каналової уваги та багатомасштабним блокам, що усуває обмеження традиційних x-vector моделей у врахуванні довготривалого контексту. Порівняно з ResNet, архітектура є обчислювально ефективнішою та менш схильною до перенавчання, зберігаючи високу точність у складних акустичних умовах. Використання ECAPA-TDNN як уніфікованої бази ознак для модулів верифікації мовця та антиспуфінгу підвищує стійкість системи до сучасних атак Deepfake та створює основу для побудови масштабованих, надійних голосових біометричних платформ.

ПЕРЕЛІК ПОСИЛАНЬ

1. Snyder D., Garcia-Romero D., Sell G., Povey D., Khudanpur S. X-Vectors: Robust DNN Embeddings for Speaker Recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Calgary, AB, Canada, 2018. P. 5329-5333. DOI : https://doi.org/10.1109/ICASSP.2018.8461375.
2. Huang H., Xiang X., Yang Y., Ma R., & Qian Y. AISPEECH-SJTU accent identification system for the Accented English Speech Recognition Challenge. *arXiv preprint arXiv:2102.09828*. 2021.
3. He K., Zhang X., Ren S., & Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2021. P. 770–778. DOI : https://doi.org/10.1109/CVPR.2016.90
4. Desplanques B., Thienpondt J., Demuynck K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. *Proc. Interspeech*, 2020. P. 3830-3834. DOI : https://doi.org/10.21437/Interspeech.2020-2650
5. Alsalihi M.H., Sztahó D. (2025). Spoof speech classification using deep speaker embeddings and machine learning models. *Array*. 2025. Vol. 27, 100494. DOI : https://doi.org/10.1016/j.array.2025.100494 .