УДК 004.934.2

К.т.н., ст. викладач Наливайчук М.В., магістрант Єгоров М. А.

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

ЗАСОБИ АДАПТИВНОГО ШУМОПОГЛИНЕННЯ АУДІОСИГНАЛІВ НА БАЗІ ЗГОРТКОВОЇ НЕЙРОННОЇ МЕРЕЖІ

Mykola V. Nalyvaichuk, senior lecturer; Yehorov Mykhail, student

Adaptive methods for audio denoising using convolutional neural networks are considered. A hybrid approach combining spectral subtraction and deep convolutional post-filtering is proposed. The method improves noise suppression quality under nonstationary conditions and reduces artifacts typical for classical spectral algorithms.

Вступ

Шумопоглинання аудіосигналів є важливим завданням у системах розпізнавання мовлення, мультимедійній обробці, телекомунікаційних системах тощо. У зв’язку зі зростанням використання голосових технологій, особливо актуальною стає задача ефективного шумопоглинання. Шум та завади у аудіозаписах суттєво знижують якість комунікації та загальне сприйняття аудіоконтенту. В умовах реального часу ця проблема стає критичною.

Попри різноманітність методів, існує проблема балансу між якістю та швидкістю роботи алгоритму. До прикладу, такі методи як спектральне віднімання або фільтрація Вінера мають високу швидкість завдяки простоті алгоритму, але часто призводять до появи артефактів у аудіосигналі. З іншого ж боку, методи на основі глибинного навчання мають високу якість очищення, а також адаптацію до різних видів шумів, однак вимагають значних ресурсів. Крім того, не завжди можуть працювати в режимі, близькому до реального часу.

Постановка задачі

Метою роботи є створення двоетапного методу адаптивного шумопоглинання, який поєднує спектральне віднімання на першому етапі та CNN-модель для доочищення на другому. Відповідно до поставленої мети необхідно розв’язати такі задачі: проаналізувати існуючі методи шумопридушення та визначити їх недоліки; розробити структуру CNN для корекції спектральних артефактів після первинного шумозаглушення;

забезпечити адаптивність алгоритму до різних типів шуму; оцінити ефективність методу на реальних аудіозаписах.

Термінологія

Спектральне віднімання – класичний метод шумопридушення, що передбачає віднімання оцінки спектра шуму від спектра поточного сигналу.

STFT (Short-Time Fourier Transform) – короткочасне перетворення Фур'є, що дозволяє представити аудіосигнал у час-частотній області.

CNN (Convolutional Neural Network) – тип нейронної мережі, ефективної для аналізу двовимірних структур, зокрема спектрограм аудіосигналів.

Аналіз існуючих підходів та алгоритмів

Фільтр Вінера. Метод базується на мінімізації середньоквадратичної помилки між очищеним та справжнім сигналом. Алгоритм ефективний для стаціонарних шумів з відомими статистичними характеристиками. Однак потребує апріорних знань про спектральну щільність потужності. Фільтр Вінера забезпечує швидку обробку. Але обмежене покращення SNR, зазвичай у межах 5-6 дБ. Основним недоліком є погіршення якості при нестаціонарних шумах, що характерні для реальних умов запису.

Спектральне віднімання. Цей метод базується на відніманні оцінки спектру шуму від спектру зашумленого сигналу в частотній області. Алгоритм складається з декількох етапів. Застосування короткочасного перетворення Фур'є (STFT), оцінка профілю шуму на основі сегментів без корисного сигналу, віднімання профілю з коефіцієнтом надвіднімання та застосування спектральної підлоги для запобігання повного обнулення. До переваг можна віднести простоту реалізації та високу швидкість обробки. Серед недоліків: схильність до появи "музичного шуму" – характерних артефактів у вигляді випадкових тональних компонент, що виникають через варіабельність шуму та помилки оцінки його профілю.

DNN-based методи (Deep Neural Networks). Глибокі нейронні мережі навчаються відображенню між зашумленими та чистими спектрограмами або безпосередньо між аудіохвилями. Архітектури варіюються від простих повнозв'язних мереж до складних рекурентних та згорткових структур. Ці методи забезпечують високу якість очищення та здатність адаптуватися до різноманітних типів шуму. А також можливість навчання на великих наборах даних. Однак вони потребують великих обчислювальних ресурсів. Особливо для рекурентних архітектур. Час обробки на CPU може перевищувати 1-2 секунди на фрагмент, що робить їх непридатними для застосувань реального часу без апаратного прискорення.

SEGAN (Speech Enhancement Generative Adversarial Network). Метод на основі генеративно-змагальних мереж, що складається з генератора з encoder-decoder архітектурою та дискримінатора, які навчаються у змагальному режимі. Генератор намагається створити очищений сигнал, невідрізнимий від справжнього чистого сигналу. А дискримінатор навчається відрізняти згенеровані сигнали від справжніх. SEGAN забезпечує високу якість очищення та природність. Недоліки включають нестабільність навчання (проблема mode collapse, коли генератор починає створювати обмежений набір виходів), складність налаштування балансу між генератором та дискримінатором. Крім того, архітектура потребує вдвічі більше пам'яті.

WaveNet. Авторегресивна модель на основі згорткових шарів з розширеною згорткою (dilated convolution), що працює безпосередньо з аудіохвилями в часовій області. Архітектура включає стеки dilated causal convolutional шарів. Це дозволяє моделі мати велике рецептивне поле при відносно невеликій кількості шарів. WaveNet демонструє найвищу якість серед розглянутих методів, відмінну натуральність звучання та здатність генерувати високоякісні аудіосигнали. Однак цей метод є найповільнішим. Складність архітектури та великий розмір моделі ускладнюють навчання, впровадження та застосування.

U-Net для аудіо. Адаптація архітектури U-Net. Спочатку була розроблена для сегментації медичних зображень, для обробки спектрограм. Архітектура складається з encoder (contracting path) та decoder (expanding path). Encoder послідовно зменшує просторову роздільність при збільшенні кількості ознакових карт, а decoder симетрично розширює представлення. Ключовою особливістю є прямі з'єднання між відповідними рівнями encoder та decoder. Це дозволяє зберігати високочастотні деталі та точну просторову локалізацію. U-Net демонструє хороший баланс між якістю та швидкістю. Але без додаткової попередньої обробки може бути недостатньо ефективною для складних шумових середовищ з високим рівнем шуму або нестаціонарними характеристиками.

Пропонований метод

Запропонований метод для шумопридушення аудіосигналів базується на двоетапній обробці, що поєднує швидкість класичних методів з якістю глибокого навчання.

1. Етап адаптивного спектрального віднімання. На першому етапі зашумлений аудіосигнал x(t) перетворюється у частотно-часову область за допомогою STFT з параметрами n_fft=2048 та hop_length=512. Для оцінки профілю шуму використовується початковий сегмент тривалістю 0.5-1.0 секунди, де припускається відсутність корисного сигналу.

Магнітудний спектр цього сегменту усереднюється по часу, формуючи оцінку $|\hat{N}(\omega)|$ для кожної частоти. Спектральне віднімання виконується за формулою: $|\hat{S}_1(\omega,m)| = \max(|X(\omega,m)| - \alpha\cdot|\hat{N}(\omega)|, \beta\cdot|X(\omega,m)|)$, де $\alpha=2.0$ – коефіцієнт надвіднімання, $\beta=0.02$ – спектральна підлога. Часовий сигнал відновлюється через iSTFT зі збереженням оригінальної фази. Результат: зменшення шуму на 70% (покращення SNR на 7 дБ), але з можливими артефактами "музичного шуму".

2. Етап CNN обробки з U-Net архітектурою. Сигнал після першого етапу перетворюється в спектрограму та нормалізується: $\tilde{S}_1 = (|\hat{S}_1| - \mu)/(\sigma + \varepsilon)$. Нормалізована спектрограма подається на вхід CNN з архітектурою U-Net. Encoder складається з чотирьох рівнів згорткових блоків (32, 64, 128, 256 фільтрів) з MaxPooling після кожного рівня. Bottleneck містить 512 фільтрів. Decoder симетрично розширює представлення через upsampling. Ключова особливість – skip connections між відповідними рівнями encoder та decoder, що зберігають високочастотні деталі мовлення. Перед конкатенацією застосовується білінійна інтерполяція для вирівнювання розмірів: якщо Dec_i.shape ≠ Enc_i.shape, то Dec_i = F.interpolate(Dec_i, size=Enc_i.shape). Це забезпечує обробку аудіо довільної довжини. Вихідна спектрограма денормалізується: $|\hat{S}_2| = \max(\tilde{S}\_CNN\cdot\sigma + \mu, 0)$, фінальний сигнал відновлюється через iSTFT.

Висновки

Запропонована двоетапна система адаптивного шумопридушення успішно поєднує обчислювальну ефективність спектрального віднімання з високою якістю CNN-обробки. Система демонструє стабільні результати для різних типів шуму та придатна для практичного застосування в телекомунікаціях, системах розпізнавання мовлення та обробці аудіо в режимі, близькому до реального часу.

Література

1. Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979.
2. Loizou P. Speech Enhancement: Theory and Practice. CRC Press, 2013.
3. Pascual S., Bonafonte A., Serrà J. SEGAN: Speech Enhancement Generative Adversarial Network. InterSpeech, 2017.
4. Rethage D., Pons J., Serra X. A WaveNet for Speech Denoising. ICASSP, 2018.
5. Luo Y., Mesgarani N. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. IEEE/ACM TASLP, 2019.