УДК 004.934.2:004.85

Д.т.н., проф. Терейковський І. А., магістрант Коротич О. С.

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

ОПИМІЗАЦІЯ НЕЙРОМЕРЕЖЕВОЇ МОДЕЛІ ВИЗНАЧИННЯ ОСНОВНОЇ ЧАСТОТИ ГОЛОСУ ЛЮДИНИ

Abstract

Igor Tereykovsky, Doctor of Technical Sciences, Prof.; Oleksandr Korotych, student
Optimization of a neural network model for determining the fundamental frequency of human speech

The thesis presents a new neural network architecture for analyzing human speech audio that drastically reduces the computational requirements compared to current solutions, while matching their quality. Experimental studies have confirmed the performance improvements of the proposed neural network architecture.

Вступ

Аналіз основної частоти (F0) людського голосу є фундаментальною задачею в обробці аудіо сигналів. Значення F0, що суб'єктивно сприймається як висота тону, несе ключову інформацію про інтонацію, емоційний стан мовця. F0 використовується в таких галузях, як розпізнавання мовлення, музичні інформаційні системи, біометрична ідентифікація та клінічна діагностика розладів голосу.

Сучасні підходи до визначення F0 все частіше спираються на глибокі нейронні мережі, які демонструють високу точність та стійкість до шумів. Однією з передових моделей у цій галузі є CREPE, яка досягає найвищої точності завдяки своїй складній згортковій архітектурі [1]. Однак висока точність CREPE досягається ціною значних обчислювальних ресурсів, що обмежує її застосування на пристроях з обмеженими можливостями, таких як мобільні телефони, вбудовані системи та веб-застосунки, що працюють у реальному часі.

Постановка задачі

Метою даного дослідження є розробка та експериментальна перевірка методу зменшення обчислювальних вимог нейромережевої моделі визначення основної частоти голосу людини, що базується на результатах моделі CREPE. Ключове завдання – створити значно "легшу" архітектуру,

яка, використовуючи принципи попередньої обробки сигналу та дистиляції знань, зможе досягти точності, порівнянної з оригінальною моделлю CREPE, при значно менших витратах на обчислення.

Термінологія

F0 (Основна частота) – це найнижча частота хвилі періодичної форми. В аналізі людського голосу F0 є фізичним відповідником акустичного поняття "висота тону" і є критично важливим параметром для характеристики просодії, інтонації та ідентифікації мовця.

CREPE – сучасна нейромережева моделі на визначення F0, яка працює безпосередньо з аудіо сигналом. Вона відома своєю високою точністю, але водночас є обчислювально складною через глибоку згорткову архітектуру.

Автокодувальник – це тип нейронної мережі, що навчається створювати ефективне, стиснене представлення вхідних даних (кодування) у латентному просторі, а потім відновлювати з цього представлення вихідний сигнал (декодування) з мінімальними втратами [2].

Віконне перетворення Фур'є – це метод аналізу сигналів, що дозволяє визначити, як змінюється частотний спектр сигналу з часом [3]. Сигнал розбивається на короткі часові проміжки (вікна), і для кожного з них виконується перетворення Фур'є. Результатом є спектрограма.

Фокальна втрата – це модифікація стандартної функції втрат бінарної перехресної ентропії, розроблена для вирішення проблеми дисбалансу класів [4]. Вона зменшує внесок у загальну втрату від "легких", правильно класифікованих прикладів (яких зазвичай більшість), і змушує модель концентруватися на "складних" прикладах. Це особливо корисно, коли один клас значно переважає інший.

Аналіз існуючих підходів та алгоритмів

Історично для визначення F0 використовувалися класичні алгоритми, такі як YIN, pYIN або RAPT, що базуються на аналізі автокореляційної функції сигналу. Ці методи є швидкими та не потребують значних ресурсів, але їхня точність суттєво знижується за наявності шумів, реверберації або при аналізі немодальних типів фонації.

З розвитком глибокого навчання з'явилися нейромережеві підходи. Модель CREPE стала стандартом у цій галузі. Її архітектура складається з шести згорткових шарів, які поступово виділяють ознаки з аудіосигналу, та фінального повно зв’язного шару, який генерує розподіл ймовірностей по 360 центрованих частотних бінах. Цей підхід дозволяє моделі самостійно навчатися оптимальним фільтрам для аналізу сигналу, що забезпечує високу точність. Проте, саме ця глибина і велика кількість параметрів у згорткових шарах роблять CREPE обчислювально "важкою". Таким чином,

існує ніша для моделей, що поєднують точність CREPE з ефективністю класичних методів.

Опис пропонованого методу

Для вирішення поставленої задачі ми пропонуємо змінити підхід до аналізу сигналу, відходячи від “чистого” навчання на сирих аудіоданих, яке використовується в CREPE. Замість того, щоб змушувати модель з нуля вивчати складні ознаки, ми інтегруємо етап попередньої обробки на основі віконного перетворення Фур’є, що дозволяє значно спростити завдання для нейронної мережі. Запропонована архітектура складається з двох послідовних модулів: кодувальника, що трансформує спектрограму в компактне латентне представлення та предиктора F0, що аналізує це латентне представлення для визначення основної частоти. Ключова оптимізація полягає у тому, як саме працює кодувальник. Ми реалізуємо його на базі архітектури автокодувальника, але його завданням є не відтворення початкового аудіо, а реконструкція амплітудної спектрограми, отриманої після віконного перетворення Фур’є. Такий підхід перетворює завдання моделі з комплексного вивчення фільтрів для аналізу сирого сигналу на значно простіше завдання - ефективну компресію вже інформативного частотного представлення. Оскільки перетворення Фур'є вже виконало основну роботу з виділення частотних складових, кодувальнику потрібна значно менша кількість згорткових шарів, що безпосередньо призводить до суттєвого зменшення обчислювальної складності.

Стиснений латентний вектор, згенерований кодувальником, подається на вхід предиктора F0, завдання якого тепер значно спрощується до аналізу цього компактного представлення. Для навчання предиктора ми використовуємо метод дистиляції знань, навчаючи його не на “правдивих” даних, а на імітуванні вихідних ймовірнісних розподілів повної моделі CREPE. Це дозволяє нашій легкій моделі успадкувати узагальнюючу здатність значно складнішої моделі. Оскільки вихід CREPE є вектором з 360 значень, де лише одиниці є значущими, ми стикаємося з проблемою сильного дисбалансу класів. Для її вирішення ми застосовуємо фокальну втрату - модифікацію бінарної перехресної ентропії, яка змушує модель концентруватися на правильному визначенні невеликої кількості активних частотних бінів, ігноруючи велику кількість неактивних, що значно підвищує ефективність навчання. Таким чином, комбінація попередньої обробки, спрощеної архітектури та спеціалізованої функції втрат дозволяє створити обчислювально ефективну модель з високою точністю.

Результати експериментальних досліджень

Запропонована оптимізація була реалізована мовою програмування Python з використанням фреймворку PyTorch. Для порівняння була взята повна версія моделі CREPE. Аналіз проводився у двох аспектах: якість відтворення результатів та обчислювальна ефективність.

Спочатку розглянимо якість результатів моделі.

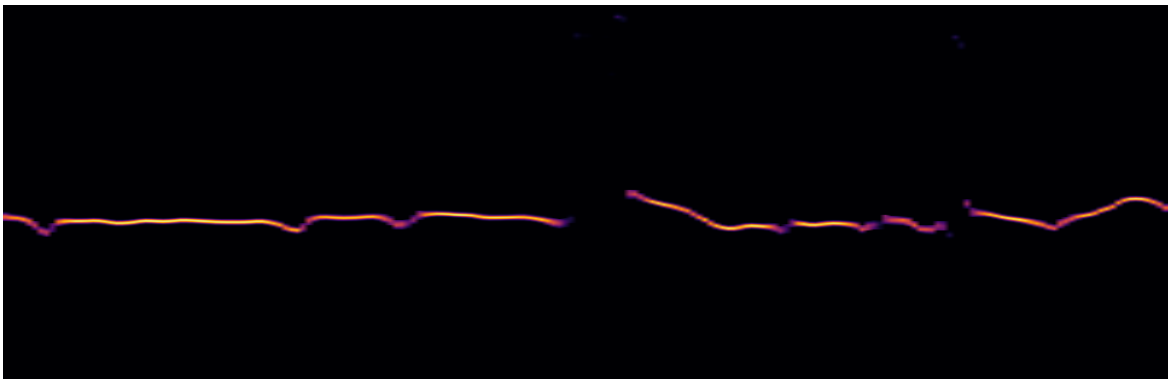

Рис. 1. Результат CREPE

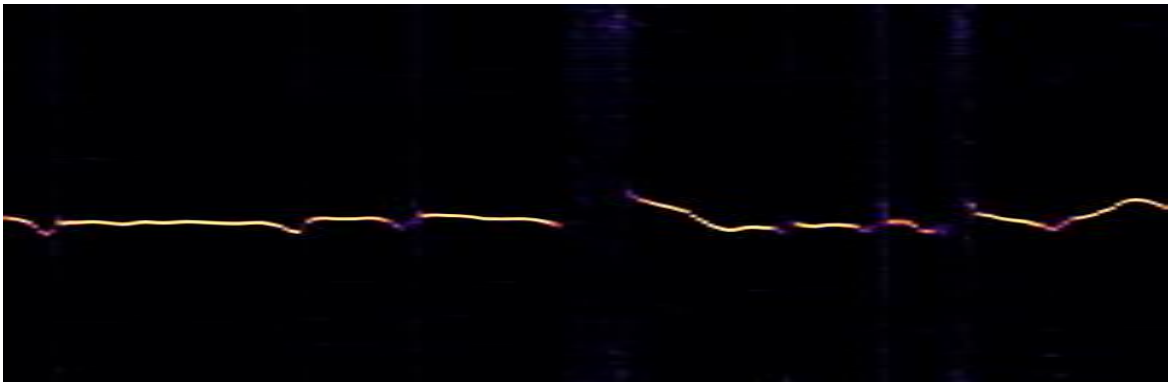

Рис. 2. Результат нашої моделі

На рисунках 1 та 2 представлено теплові карти виходів, де по горизонтальній осі відкладено час, по вертикальній осі – частотні біни, а інтенсивність кольору відображає впевненість моделі у відповідній частоті для кожного моменту часу.

На Рис. 3 для кожного часового кроку обирався бін з максимальною ймовірністю. Якщо ця ймовірність була нижчою за поріг 50%, значення F0 вважалося нульовим, що відповідає неозвученому сегменту.

Аналіз показує, що запропонована модель майже ідеально повторює результати CREPE. Незначні відхилення помітні у високочастотних бінах під час пауз: наша модель іноді показує залишкову низьку впевненість, тоді як у CREPE вона повністю відсутня. Це, ймовірно, є наслідком недостатньої репрезентації прикладів з високим тоном F0 у навчальних даних.

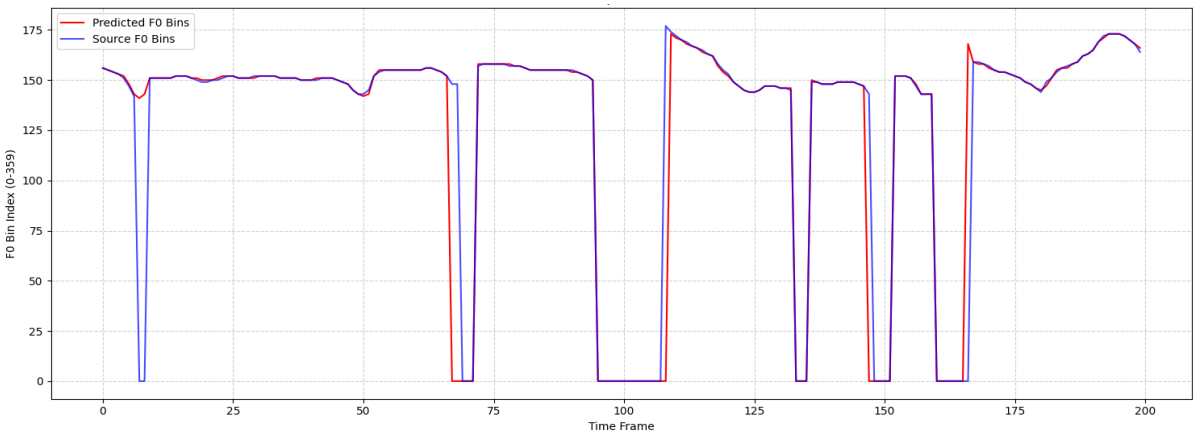

Рис. 3 Графік накладених результатів

Для оцінки обчислювальної ефективності було проведено тестування у двох сценаріях: обробка одного аудіофайлу тривалістю 16 секунд та пакетна обробка восьми 8-секундних файлів. Результати, усереднені по 30 запусках для стабільності, наведені в Табл. 1.

Таблиця 1

Результати

Сценарій	Модель	Сер. час виконання, с (станд. відх.)	Пікове GPU, МБ
Один файл (1x16 с)	CREPE	0.1662 (±0.0008)	968.32
	Наша модель	0.0090 (±0.0001)	402.34
	Покращення	*~18.5 разів швидше*	*~2.4 рази менше*
Пакет файлів (8x8 с)	CREPE	0.6728 (±0.0024)	2847.33
	Наша модель	0.0346 (±0.0007)	487.83
	Покращення	*~19.4 разів швидше*	*~5.8 разів менше*

Дані тестування демонструють значне зменшення обчислювальних вимог. Запропонована модель виконує обробку більш ніж у 18.5-19.4 разів швидше та вимагає в 2.4-5.8 рази менше відеопам'яті у порівнянні з оригінальною архітектурою CREPE, при цьому демонструючи високу стабільність результатів.

Висновки

У ході даного дослідження було розроблено та успішно протестовано метод для суттєвого зменшення обчислювальних вимог нейромережевої моделі визначення основної частоти голосу. Запропонований підхід, що поєднує попередню обробку сигналу за допомогою віконного перетворення

Фур'є, стиснення ознак автокодувальником та дистиляцію знань від моделі CREPE, довів свою високу ефективність.

Експериментально встановлено, що розроблена модель досягає точності, яка є практично ідентичною до еталонної моделі CREPE, при цьому демонструючи значне підвищення продуктивності. Швидкість обробки аудіо зросла більш ніж у 19 разів, а споживання відеопам'яті зменшилося приблизно у 2.5 рази для маленької кількості даних і 5.8 разів для великої кількості. Це повністю вирішує поставлену задачу і відкриває можливість для інтеграції високоточного аналізу F0 в додатки реального часу та на пристрої з обмеженими ресурсами.

Література

1. Kim J. W. CREPE: A Convolutional Representation for Pitch Estimation / J. W. Kim, J. Salamon, P. Li, J. P. Bello // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2018. – P. 161–165.
2. Bank D. Autoencoders / D. Bank, N. Koenigstein, R. Giryes // arXiv preprint arXiv:2003.05991. – 2020. – 29 p.
3. Harris F. J. On the use of windows for harmonic analysis with the discrete Fourier transform / F. J. Harris // Proceedings of the IEEE. – 1978. – Vol. 66, № 1. – P. 51–83.
4. Lin T.-Y. Focal Loss for Dense Object Detection / T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár // Proceedings of the IEEE International Conference on Computer Vision (ICCV). – 2017. – P. 2980–2988.