D25-001

應用於腦機介面之高能效意念至文字轉換處理器晶片

An Energy-Efficient Neural Signal Processor in Speech Decoding for Brain-Machine Interface

隊伍名稱｜要記得帶綠色乖乖 Remember to Bring Green Kuai Kuai

隊　　長｜張惇宥 / 臺灣大學電子工程學研究所

隊　　員｜王政邦 / 臺灣大學電子工程學研究所

指導教授

楊家驤｜臺灣大學電機工程學系暨電子工程學研究所

美國加州大學洛杉磯分校電機博士，現為臺灣大學電機工程學系暨電子工程學研究所教授。目前擔任IEEE頂尖期刊JSSC副主編，曾擔任頂尖國際會議（ISSCC、VLSI Symposium）技術議程委員。曾獲ISSCC傑出技術論文獎、ISSCC遠東區最佳論文獎、國科會傑出研究獎、吳大猷先生紀念獎、中國電機工程學會傑出電機教授獎、胡正明半導體創新獎、傑出人才發展基金會年輕學者創新獎、臺灣積體電路設計學會傑出年輕學者獎等獎項。

研究領域

AI晶片設計、基頻通訊積體電路、生醫訊號處理晶片設計。

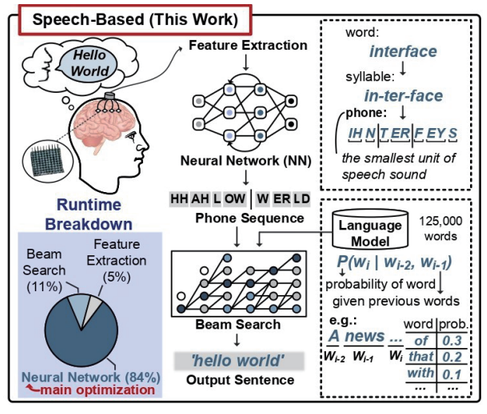
作品摘要

腦機介面能在大腦與外部機器之間建立直接的通訊橋樑，在未來的應用領域極為廣泛，包括擴增實境/虛擬實境介面、神經義肢、以及機器控制等，能顯著改善人機互動體驗。目前已有多種型態的腦機介面系統，包括包含視覺刺激型、手寫型與語音型等。視覺刺激型依賴使用者觀看閃爍目標來選擇輸入內容，手寫型則透過解析使用者想像書寫軌跡來推測文字，兩者在實用性與速度上仍有侷限。而語音型腦機介面能夠將使用者試圖說話的意圖直接轉換為文字，具備最高的通訊速率與自然性，在這類型中，神經網路會根據大腦訊號所提取出的特徵，推論使用者可能發出的語音音素（最小語音單位）的機率，並配合語言模型與波束搜尋演算法來解碼最可能的文字序列。過去已有針對視覺刺激型與手寫型腦機介面開發專用處理器，然而，由於語音型腦機介面涉及複雜的神經網路運算與大規模語言模型推理，其對運算資源與能效的要求極高，至今仍少有專用處理器針對此需求進行設計與優化。

為解決上述挑戰，我們提出一款用於即時語音型腦機介面的高能效神經訊號處理器，具備最高每分鐘200字的對話語音解碼速率。此晶片整合了說話嘗試檢測器、特徵提取器、神經網路引擎以及波束搜尋引擎等核心模組，有以下的技術特點：

1. 通道選擇技術：將說話嘗試檢測器的通道數目從128減少到16，在使用者未嘗試說話時，能耗降低了46%。
2. 權重編碼機制：結合稀疏編碼與混合精度運算，大幅降低神經網路模型的記憶體需求高達80%；在處理單元陣列中採用混合精度乘法器設計，相較傳統全精度運算，硬體面積縮小27%。
3. 神經網路運算優化：透過運算重排序技術減少55%延遲，並導入「部分和快取」機制重用中間運算結果，有效降低25%的總運算量。再配合輸入與權重稀疏性的利用，進一步將運算延遲壓縮高達95%。
4. 近似Top-k選擇器：於波束搜尋階段採用近似排序架構，將傳統排序所需比較器數量降低至1/16，兼顧硬體簡化與準確度。

本設計採用40奈米製程，在說話解碼性能上達到了16.6%的音素錯誤率和23.5%的詞錯誤率，與現有最先進設計相比，實現了16.7到42.6倍的解碼速度提升。



圖一 語音型腦機介面。

Abstract

Brain-Machine Interface (BMI) technology establishes a direct communication link between the human brain and external devices, offering tremendous potential in future applications such as augmented/virtual reality (AR/VR) interfaces, neuroprosthetics, and machine control. These systems can significantly enhance human-computer interaction. Current BMI implementations are categorized into visual-stimulus-based, handwriting-based, and speech-based types. Visual-stimulus BMIs rely on the user's gaze to select flashing targets, while handwriting BMIs decode imagined writing trajectories into text. However, both approaches have limitations in practicality and communication speed. In contrast, speech-based BMIs offer the most natural interaction and the highest communication throughput by decoding the user's intent to speak directly into text. These systems utilize neural networks to infer the probability distribution of phonemes—based on features extracted from neural signals—and use language models and beam search algorithms to generate the most likely word sequences. While processors have been developed for visual and handwriting BMIs, speech-based BMIs demand far greater computational resources and energy efficiency due to their reliance on complex neural network computations and large-vocabulary language model inference. To date, few dedicated hardware solutions have been proposed to address this challenge.

To overcome these barriers, we propose a real-time, energy-efficient neural signal processor tailored for speech-based BMI applications, capable of decoding up to 200 words per minute. The proposed chip integrates four major modules: a speech-attempt detector, a feature extractor, a neural network engine, and a beam search engine. Key innovations include:

1. Channel Selection: Reducing the number of channels for speech-attempt detection from 128 to 16, lowering power consumption by 46% during idle periods.
2. Weight Encoding: Combining sparse encoding with mixed-precision computation, this method reduces neural network memory requirements by up to 80%. A custom mixed-precision multiplier array also reduces chip area by 27% compared to full-precision designs.
3. Neural Network Optimization: Introducing computation reordering reduces inference latency by 55%, while partial-sum caching lowers total operations by 25%. Combined with input and weight sparsity, overall latency is reduced by up to 95%.
4. Approximate Top-k Selector: In the beam search stage, an approximate sorting architecture reduces the number of required comparators to 1/16 of conventional designs while maintaining decoding accuracy.

Fabricated using a 40nm CMOS process, the chip achieves a phoneme error rate of 16.6% and a word error rate of 23.5%. Compared to the state-of-the-art, the proposed design delivers a 16.7× to 42.6× improvement in decoding speed.

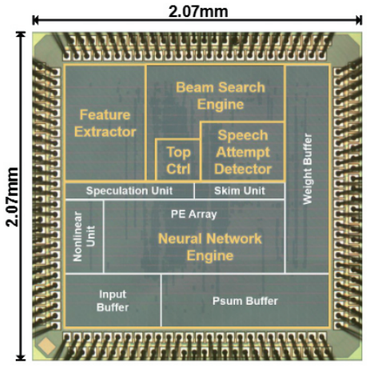


Fig. 2 Chip micrograph.